THE USE AND INTERPRETATION OF HIGH SCHOOL TESTS

LONGMANS' EDUCATION SERIES

FREDERIC B. KNIGHT, *General Editor*

FOUNDATIONS OF HUMAN NATURE
By John M. Dorsey

PROBLEMS OF THE TEACHING PERSONNEL
By Dennis H. Cooke

INTRODUCING EDUCATION
By J. Herbert Blackhurst

WORKBOOK FOR INTRODUCING EDUCATION
By J. Herbert Blackhurst

THE PSYCHOLOGY OF ABNORMAL PEOPLE
By John J. B. Morgan

WORKBOOK IN ABNORMAL PSYCHOLOGY [*Forms A & B*]
By John J. B. Morgan

WORKBOOK IN EDUCATIONAL MEASUREMENTS [*Forms A & B*]
By Harry A. Greene

THE USE AND INTERPRETATION OF ELEMENTARY SCHOOL TESTS
By Harry A. Greene and *A. N. Jorgensen*

THE USE AND INTERPRETATION OF HIGH SCHOOL TESTS
By Harry A. Greene and A. N. Jorgensen

STUDY MANUAL IN ELEMENTARY STATISTICS
By Everett F. Lindquist and *George D. Stoddard*

OBJECTIVES AND PROCEDURES IN CIVIC EDUCATION
By Charles C. Peters

THE TEACHING OF HIGH SCHOOL ENGLISH
By Virginia J. Craig

MAKING A HIGH SCHOOL SCHEDULE OF RECITATIONS
By R. C. Puckett

IN PREPARATION

PSYCHOLOGY OF HIGH SCHOOL SUBJECTS
By H. H. Remmers

NEURO-PHYSIOLOGY
By Lee Edward Travis, Waid Wright Tuttle,
and *Frederic B. Knight*

THE USE AND INTERPRETATION OF HIGH SCHOOL TESTS

BY

HARRY A. GREENE, Ph.D.

ASSOCIATE PROFESSOR OF EDUCATION, AND
DIRECTOR OF BUREAU OF EDUCATIONAL RESEARCH AND SERVICE
UNIVERSITY OF IOWA

and

ALBERT N. JORGENSEN, Ph.D.

PRESIDENT OF CONNECTICUT STATE COLLEGE
FORMERLY PROFESSOR OF EDUCATION
UNIVERSITY OF BUFFALO

LONGMANS, GREEN AND CO.

NEW YORK • LONDON • TORONTO

1936

LONGMANS, GREEN AND CO.
114 FIFTH AVENUE, NEW YORK
221 EAST 20TH STREET, CHICAGO
88 TREMONT STREET, BOSTON

LONGMANS, GREEN AND CO. LTD.
39 PATERNOSTER ROW, LONDON, E C 4
6 OLD COURT HOUSE STREET, CALCUTTA
53 NICOL ROAD, BOMBAY
36A MOUNT ROAD, MADRAS

LONGMANS, GREEN AND CO.
480 UNIVERSITY AVENUE, TORONTO

GREENE AND JORGENSEN
THE USE AND INTERPRETATION
OF HIGH SCHOOL TESTS

COPYRIGHT · 1936
BY LONGMANS, GREEN AND CO.

ALL RIGHTS RESERVED, INCLUDING THE
RIGHT TO REPRODUCE THIS BOOK, OR
ANY PORTION THEREOF, IN ANY FORM

FIRST EDITION

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

THIS book is a practical introductory discussion of the essential principles of measurement in education. It is designed especially for the high school teacher and the student of secondary education. This volume itself is a rather complete revision and expansion of an earlier treatment which appeared under the title of *The Use and Interpretation of Educational Tests.* Certain inadequacies in the presentation of illustrative material suitable for both elementary and secondary school teachers which were apparent in the earlier volume are overcome in this revision by the preparation of two separate treatments.

This volume is concerned with the problems of measurement of instruction encountered in the junior and senior high school grades. Wherever possible the illustrations are taken from the material designed for use in these grades. The other volume is just as specifically addressed to those students and teachers who primarily face the problems of instruction and measurement as they exist at the elementary school level. Naturally, many of the basic measurement problems are common to both the elementary and the secondary levels. However, it is believed that the specific elaboration and illustration of these problems in terms of materials taken almost solely from the teacher's own fields of thought and interest distinctly improve the value of this volume as a textbook in educational measurements in the hands of students and teachers.

This revised treatment is planned to provide a systematic handbook for any person who feels the need of a practical and easily understood discussion of the fundamental ideas and tools of measurement in secondary education. It is written from the point of view of the classroom teacher, and in non-technical language. Its direct method of attack,

the practical classroom point of view taken throughout these pages, and the concrete illustrations taken from the students' own immediate fields of knowledge and interest are methods taken to insure a strong appeal to the group for which it is intended. Only through the active enlistment of the interest of the classroom teacher and of the student who himself is to serve in that capacity will the fullest possibilities of measurement in education ever come to be realized.

The authors believe that the chief purpose to be served by educational tests in the classroom is the improvement of instruction and of learning through the collection of more detailed and more accurate information concerning the existence and causes of individual pupil difficulties in learning. This implies the wider use of more exact methods of measuring pupil accomplishment, of locating, diagnosing, and remedying pupil weaknesses. The authors believe also that the responsibility for the use of these instruments for the accurate measurement of accomplishment and for the location, diagnosis, and remedy of pupil difficulties must be more and more assumed by the classroom teacher. This does not mean any less responsibility on the part of the principal or supervisor. It means a clearer recognition of it on the part of the classroom teacher, for after all the use and interpretation of educational tests is a co-operative project. The goal of all good supervision is the improvement of the conditions under which teachers teach and children learn.

To accomplish these purposes, much more emphasis must be placed upon the training of the prospective teacher, as well as of the more experienced teacher, in the construction and use of measuring devices and in the development and application of remedial materials following diagnosis. This book places in the hands of the normal school and teachers college student, the classroom teacher, and the supervisor of the junior and senior high school grades basic suggestions of ways in which educational measuring instruments may be of practical use, and general rules for guidance in constructing, selecting, using, and interpreting educational tests as valuable instructional aids. The critical study of this discussion, the thoughtful preparation of the discussion exer-

cises, and the working of the practical problems which form an integral part of the book, should bring the teacher or student to a better appreciation of the possibilities of educational tests as teaching and supervisory instruments.

The indebtedness of the authors to the many experienced teachers and supervisors who have contributed indirectly as well as directly to the formulation of much of the material incorporated herein is gratefully acknowledged. The authors' special thanks are also due to the many users of the earlier text who made many helpful suggestions as to its reorganization and revision. They are also grateful to the numerous authors and publishers who have given permission for the reproduction of many selected portions of their work and publications. Specific acknowledgments of these are made throughout the text.

Special acknowledgments are due Professor Frederic B. Knight for his careful study of this manuscript and for his many valuable and constructive suggestions as to its organization and content.

EDITOR'S INTRODUCTION

The authors of this volume have had close contact and extensive experience with the teaching of educational measurements to elementary and secondary school teachers as well as with the practical problems of the improvement of classroom instruction. Such experience qualifies them in a happy way to understand not only Educational Tests themselves but also the needs of teachers and supervisors. A union of this practical experience and technical training possessed by the authors produced a book on measurements which was sound in theory and effective in presentation. This book was entitled *The Use and Interpretation of Educational Tests* and for five years enjoyed the confidence of a large number of college teachers. It has been the basic text for many thousands of students and teachers. It is now time to offer a revision.

The present volume is an enlarged and expanded treatment of the subject rather than a reprinting of its progenitor. While it is not found necessary to contradict any fundamental position taken in the older edition, several important modifications in emphasis, attention to newer tests, modifications due to the growth in the science of measurement have caused significant changes.

The classroom teacher and the busy supervisor, and those in training for teaching or supervision will find this book a carefully written fundamental text on the principles of measurement in education. Its main contribution to the growing literature on measurement is, I think, not so much in new points of view or advances in the technicalities of test construction but rather in its plainness of exposition and balance in treatment of many points which to some seem over technical. It should be thought of as a first book in measurement for those who do not at the time of studying

it know very much, if anything, about measurement in education and its application to the problems of improving classroom instruction.

A careful examination of the book will support the opinion that this book is not only excellent in content, as such, but is an effective learning instrument for students. A present hopeful tendency of collegiate teaching is to stress factors of presentation as well as problems of content. Hence a liberal supply of practice problems and exercises using critical judgment is an essential part of this book. Frequently, the instructor may deem it wise to supply occasions for the more actual "doing of the job itself" than exercises in a single volume can conveniently provide to meet this need.

Professor Greene has prepared two alternate forms of a *Work-Book in Educational Measurements*. The present book together with the Work-Book supplies to the college teacher and his students co-ordinated learning units in the field of educational measurements. Not only is the student taught, but he is given sufficient and carefully graded practice in measurement, so that upon mastery of the books he can face the problems of measurement in actual school situations with considerable confidence in his preparation to solve such problems with success.

Teachers must have an understanding of the use and interpretation of educational tests. This book is definitely directed to meet in a frank and reasonable way the need for such training. It is a basic text for classroom use in a course in educational tests and their uses. It is also well adapted for use in connection with extension classes and correspondence courses.

F. B. KNIGHT

CONTENTS

CHAPTER PAGE

I. INTRODUCTION 1

I. BACKGROUND OF THE GROWTH OF THE MEASUREMENT IDEA: 1. Educational Measurements Not a New Idea. 2. Real Pioneers in Measurements. 3. Factors in Growth of Measurement Idea. 4. Rapid Growth of Interest in Tests. 5. Purpose of This Book.

II. THE MEANING OF MEASUREMENT IN EDUCATION 9

I. WHAT TESTS ARE: 6. Measurement in Education. 7. Characteristics of Educational Tests. 8. What Tests Do. 9. What Tests Do Not Do.

II. KINDS OF TESTS: 10. Essay-Type Tests. 11. Objective Tests. 12. Tests and Scales. 13. Standardized *vs.* Informal Objective Tests. 14. Mental Tests. 15. Aptitude and Prognostic Tests. 16. Tests of Educational Achievement. 17. Quality and Product Scales. 18. Survey *vs.* Diagnostic Tests.

III. THE IMPROVEMENT OF TEACHER-MADE MEASURES OF ACHIEVEMENT 40

I. EXAMINATIONS AND THE CLASSROOM TEACHER: 19. Teacher Evaluation of Instruction. 20. Shift from Oral to Written Examinations. 21. Influence of Horace Mann.

II. FUNCTIONS OF CLASSROOM MEASUREMENT: 22. Clarification of the Aims and Purposes of Examinations. 23. Significance Attached to Examinations.

III. IMPROVEMENT OF THE TEACHER'S CLASSROOM TESTING TECHNIQUES: 24. Need for Improvement in Classroom Measurement. 25. Improving the Teacher-Made Essay Examination. 26. Specific Suggestions for Increasing the Objectivity of Scoring Essay-Type Examinations.

CHAPTER PAGE

IV. CONSTRUCTING AND USING INFORMAL OBJECTIVE EXAMINATIONS IN THE CLASSROOM 58

I. CHARACTERISTICS OF INFORMAL OBJECTIVE EXAMINATIONS: 27. Introduction. 28. The New Examination. 29. Standardized *vs.* Non-Standardized Objective Tests.

II. ADVANTAGES AND DISADVANTAGES OF THE INFORMAL OBJECTIVE EXAMINATION: 30. Advantages of the Informal Objective Examination. 31. Possible Disadvantages of Objective Examinations.

III. PRINCIPLES OF CONSTRUCTING AND USING OBJECTIVE EXERCISES: 32. Need for Specific Directions. 33. Number of Exercises to Use. 34. Selection of Subject-Matter Items to be Tested. 35. Clear Statement of Exercises. 36. Range of Difficulty of the Exercises. 37. Arrangement of the Exercises. 38. Arrangement of Answers for Convenience in Scoring. 39. Arrangement of Items According to Difficulty. 40. Elimination of Faulty Exercises. 41. Use of Various Types of Exercises. 42. Sources of Objective Test Material.

IV. PRACTICAL USES OF OBJECTIVE TEST RESULTS: 43. Instructional Uses of Objective Tests. 44. Use of Objective Test Scores as Basis for Class Grades.

V. CLASSIFICATION AND ILLUSTRATIONS OF TYPES OF OBJECTIVE EXERCISES: 45. Classification of Objective Tests. 46. Samples of Types of Objective Test Exercises.

VI. CHANCE FACTORS IN OBJECTIVE TESTS: 47. Correcting for Guessing.

V. THE USES OF STANDARDIZED EDUCATIONAL TESTS 90

48. Tests in Relation to Classroom Instruction.

I. INSTRUCTIONAL USES OF ACHIEVEMENT TESTS: 49. For Purposes of Comparison. 50. For Measuring the Efficiency of Learning. 51. For Purposes of Class Analysis and Diagnosis. 52. For Purposes of Individual Pupil Diagnosis. 53. For Pupil Gradation. 54. For Pupil Guidance.

II. USES OF MENTAL TESTS: 55. Mental Tests in the Classroom.

VI. USING STANDARD TESTS IN THE CLASSROOM 100

I. GIVING THE TESTS: 56. The Teacher's Responsibility. 57. Scoring and Interpreting the Tests.

CHAPTER PAGE

II. When to Give the Tests: 58. Periodical *vs.* Continuous Use of Tests.

III. The Basis for the Testing Program: 59. Clear-Cut Teaching Problem as Basis. 60. Steps in a Testing Program. 61. Illustrations of Problems.

IV. The Criteria for Selecting Tests: 62. Validity. 63. Reliability. 64. Objectivity. 65. Administrability. 66. Economy. 67. Adequacy of Test in Measuring Specific Functions.

V. The Critical Selection of Tests: 68. Test Rating Scales.

VI. General Rules for Administration: 69. Preparation for Administering the Test. 70. Timing. 71. Scoring.

VII. The Construction of Standardized Tests 127

I. The Process of Test Standardization: 72. Meaning of Standardization. 73. Standardized *vs.* Informal Objective Tests.

II. Validation of Content: 74. Selection of Content. 75. Determining Validity.

III. Criteria for Construction of Test Exercises: 76. Form of Exercise as Conditioning Objectivity. 77. Criteria for Selection of Exercises.

IV. Construction of Equal Forms: 78. Equating the Forms of the Test.

V. Critical Aspects of Reliability: 79. Reliability of Measurement. 80. Reliability of the Norms.

VIII. Summarizing the Results of Testing 149

I. Classification and Tabulation of Test Scores: 81. Need for a Method of Grouping Data. 82. Steps Involved in Preparing a Frequency Table. 83. Determining the Range of Scores. 84. Determining the Number and Size of Class-Intervals. 85. Determining the Limits of Class-Intervals. 86. Tabulating the Scores. 87. Summary of Steps in Making a Frequency Table.

II. Measures of Central Tendency: 88. Computing the Arithmetic Mean — Ungrouped Data. 89. Computing the Arithmetic Mean — Data Grouped in Frequency Distribution. 90. Summary of Steps in Finding Arithmetic Mean. 91. Classroom Use of the Arithmetic Mean. 92. The Median as a Measure of Central Tendency. 93. Computing a

CHAPTER PAGE

Counting Median or Mid-Measure. 94. Computing the Median from Frequency Distribution.

III. Measures of Variability: 95. Need for Measures of Variability. 96. Range as an Expression of Variability. 97. "Q" or Quartile Deviation as a Measure of Variability. 98. Computing the Quartile Deviation. 99. Standard Deviation as a Measure of Variability. 100. Computing the Standard Deviation from Ungrouped Data. 101. Computation of Sigma from Grouped Data. 102. Practical Uses of the Standard Deviation.

IV. The Relationship of Test Scores: 103. Need for Measures of Relationship. 104. The Coefficient of Correlation. 105. Tabulating Test Scores in Double Entry or Correlation Tables. 106. Steps in the Computation of the Pearson Product-Moment Coefficient of Correlation. 107. Illustration of the Computation of the Pearson Correlation Coefficient. 108. Meaning of the Correlation Coefficient. 109. Practical Uses of Correlation Method.

V. Assignment of Relative and Absolute Ranks: 110. Need for Method of Ranking. 111. Assignment of Relative Ranks. 112. Assignment of Absolute Ranks. 113. Computing the Percentiles in a Frequency Distribution.

VI. Summary.

IX. Interpreting the Results of Testing 209

I. Test Scores: 114. Meaning of a Test Score. 115. Giving Meaning to Informal Test Scores.

II. Test Norms and Standards: 116. Meaning of Standardization. 117. Test Norms. 118. Norms and Standards. 119. Uses of Test Norms and Standards. 120. Kinds of Norms. 121. Derivation of Norms. 122. Meaning of Test Norms. 123. Grade Norms. 124. Age Norms. 125. Interpretation of Grade and Age Norms.

III. Types of Derived Scores: 126. Function of Derived Scores. 127. Grade Norms as Derived Scores — Type 1. 128. Age Scores as Derived Scores — Type 2. 129. Percentile Ranks as Derived Scores — Type 3. 130. Quotients as Derived Scores — Type 4.

IV. Summary.

CHAPTER PAGE

X. THE PLACE OF TESTS IN DIAGNOSIS AND REMEDIAL TEACHING 247

I. THE PLACE OF DIAGNOSIS AND ANALYSIS: 131. The Meaning of Diagnosis. 132. Diagnosis in Education. 133. Analysis as the Basis of Diagnosis. 134. Specific Nature of Diagnosis. 135. Importance of the Diagnostic Interpretation of Test Results.

II. THE PLACE OF REMEDIAL INSTRUCTION: 136. Exact Diagnosis the Basis for Remedial Work. 137. Diagnosis as the Basis for Preventive Work. 138. Necessity for Valid Drill for Each Identified Skill.

III. SUMMARY.

XI. MENTAL TESTS AS DIAGNOSTIC AIDS 259

I. THE MEANING OF MENTAL ABILITY: 139. The Nature of Mental Ability.

II. THE MEASUREMENT OF GENERAL MENTAL ABILITY: 140. The Meaning of a Mental Test Score. 141. The Nature of the Mental Test.

III. THE KINDS OF MENTAL TESTS: 142. Individual Mental Examinations. 143. Group Tests of Mental Ability.

IV. EDUCATIONAL USES OF MENTAL TESTS: 144. For Class Diagnosis. 145. For Individual Diagnosis.

V. INTER-DEPENDENCE OF MENTAL AND ACHIEVEMENT TESTS: 146. Care in Use of Mental Test Results. 147. Inter-dependence of Achievement and Mental Test Results.

VI. DERIVED RESULTS OF MENTAL TESTING: 148. Mental Age as a Derived Score. 149. Intelligence Quotients as Derived Measures. 150. The Personal Constant as a Derived Measure.

VII. SUMMARY.

XII. MEASUREMENTS AND THE TOTAL ADJUSTMENT OF THE PUPIL 282

I. EDUCATION AS ADJUSTMENT: 151. Adjustment and Guidance. 152. School Responsibility for Adjustment. 153. Adjustment as Related to Personality, Character and Individuality.

II. GUIDANCE AS ADJUSTMENT: 154. Guidance, Vocational and Educational. 155. The Adjusted Pupil. 156. Basic Approaches to Pupil Adjustment. 157. Sources of Guidance Data.

CHAPTER | PAGE

III. Measurement of Personality and Character Traits: 158. Rating Scales for Personality and Character Traits.
IV. Measurement of Vocational Interests: 159. Strong's Vocational Interest Blank.
V. Measurement of Special Aptitudes: 160. Mechanical Aptitude. 161. Artistic Talent. 162. Aptitude in High School Mathematics.

XIII. Measurement in English and Related Subjects 311
I. Aims and Outcomes of Instruction in English: 163. The Educational and Social Importance of English. 164. Specific Objectives of English Instruction.
II. Measurement of Oral English: 165. Diagnosis of Oral English Disabilities. 166. Testing for Speech Defects.
III. Measurement of Written English Abilities: 167. Measures of General Merit of Written Composition. 168. Tests of Grammar and Usage. 169. Form and Usage Tests. 170. Analytical Measurement of Language Abilities. 171. English Diction Test.
IV. Spelling as a Factor in English Achievement: 172. Measurable Qualities in Spelling. 173. Construction of Spelling Tests. 174. Diagnosis of Spelling Disabilities.
V. Handwriting as a Factor in Language Ability: 175. Measurable Qualities in Handwriting. 176. Methods of Measurement. 177. Accuracy in Measurement of Handwriting. 178. Merit Scales. 179. Diagnosis of Handwriting Faults. 180. Remedial Work in Handwriting.

XIV. Measurement in Reading and Literature 349
I. Educational and Social Importance of Reading: 181. Reading as a Tool Skill. 182. Reading and Literature. 183. Major Objectives in Reading Instruction. 184. Measurement of Work-Study Type of Reading. 185. Remedial Drills for Work-Study Types of Reading Difficulties. 186. Other Remedial Materials in Reading. 187. Vocabulary Tests.
II. Measurement in Literature: 188. Objective Tests in Literature. 189. Reading Scales in Literature.

CHAPTER PAGE

XV. MEASUREMENT IN THE SOCIAL SCIENCES . 369

I. IMPORTANCE OF THE SOCIAL SCIENCES: 190. Objectives of Instruction in the Social Studies.

II. POSSIBILITIES OF MEASUREMENT IN THE SOCIAL SCIENCES: 191. Acquisitive Skills. 192. Measurement of Outcomes. 193. Kinds of Tests in the Social Sciences.

III. TYPES OF TESTS IN THE SOCIAL SCIENCES: 194. Measurement of Achievement in American History. 195. Textbook Tests and Study Guides in American History. 196. Measurement of Achievement in European and World History. 197. Tests for Civics and Government. 198. Geography Tests. 199. Miscellaneous Social Science Tests. 200. Diagnosis and Remedy in the Social Studies.

XVI. MEASUREMENT IN MATHEMATICS 389

I. GENERAL SIGNIFICANCE OF MATHEMATICS: 201. Significance of Mathematics. 202. Mathematics in the Junior-Senior High School.

II. ARITHMETIC IN THE JUNIOR-SENIOR HIGH SCHOOL: 203. Fundamentals of Arithmetic. 204. Survey and Diagnosis in Arithmetic.

III. MEASUREMENT IN ALGEBRA: 205. Objectives of Instruction in Algebra. 206. Measurable Qualities in Algebra. 207. Tests in First Year Algebra. 208. College Placement and Prognosis Tests.

IV. PLANE GEOMETRY: 209. Objectives of Instruction in Plane Geometry. 210. Possibilities of Measurement. 211. Measures of Achievement in Plane Geometry. 212. Instructional Tests in Plane Geometry. 213. Prognosis in Geometry.

V. SOLID GEOMETRY AND TRIGONOMETRY: 214. Solid Geometry. 215. Trigonometry.

XVII. MEASUREMENT IN THE SCIENCES 417

I. GENERAL SIGNIFICANCE OF THE SCIENCES: 216. Science and Human Progress. 217. Instructional Outcomes in the Sciences. 218. Scope of Instruction in the Sciences.

II. POSSIBILITIES AND LIMITATIONS OF MEASUREMENT: 219. Measurable Qualities. 220. Measurement in the Elementary Sciences.

III. MEASUREMENT IN GENERAL SCIENCE: 221. Objectives in General Science. 222. Types of General Science Tests.

CHAPTER PAGE

IV. Measurement in Physics: 223. Objectives in Physics. 224. Measurable Qualities in Physics. 225. Types of Physics Tests. 226. Miscellaneous Physics Tests.

V. Measurement in Chemistry: 227. Objectives in Chemistry. 228. Measurable Qualities in Chemistry. 229. Types of Chemistry Tests.

VI. Measurement in Biology: 230. Instructional Aims in Biology. 231. Types of Biology Tests.

VII. Miscellaneous Types of Science Tests: 232. Informal Objective Tests in Science. 233. American Council Co-operative Tests. 234. Iowa Every-Pupil Tests in Science.

VIII. Diagnostic and Remedial Work in the Sciences: 235. Diagnosis and Remedy.

XVIII. Measurement in the Foreign Languages 445

I. Educational Importance of the Foreign Languages: 236. Shifting Objectives of Instruction.

II. Measurement of Latin: 237. Possibilities of Measurement. 238. Achievement Tests in Latin. 239. Prognosis in Latin. 240. Remedial Instruction in Latin.

III. Measurement in the Modern Languages: 241. Measurable Elements in the Modern Languages. 242. Achievement Tests in French. 243. Achievement Tests in German. 244. Achievement Tests in Spanish. 245. Prognosis in the Modern Languages. 246. Remedial Work in Modern Languages. 247. Miscellaneous Language Tests.

XIX. Measurement in Fine Arts . . 471

I. Measurable Qualities in Music: 248. Aims and Outcomes in Musical Education.

II. Measurement of Musical Abilities and Achievement: 249. Basic Musical Talent. 250. Musical Achievement. 251. Remedial Materials in Music.

III. Measurement of Art Abilities and Achievement: 252. General Outcomes of Art Instruction. 253. Specific Outcomes. 254. Measurement of Artistic Talent. 255. Measurement of Artistic Accomplishment. 256. Applied Art.

XX. Health and Physical Education . 493

I. Health and Physical Education in the Educational Program: 257. Educational Significance

CHAPTER PAGE

of Health Education. 258. Objectives of Instruction in Health. 259. Objectives of Physical Education. 260. Measurable Qualities in Health and Physical Education.

II. MEASUREMENT OF INFORMATION ON HEALTH: 261. Health Knowledge Tests.

III. MEASUREMENT OF PHYSICAL GROWTH AND STRENGTH: 262. Weight-Height-Age Tables. 263. Physical Capacity Tests. 264. Scales of Motor Ability. 265. Physical Achievement Tests.

IV. PREVENTIVE AND CORRECTIVE WORK IN HEALTH AND PHYSICAL EDUCATION: 266. Prevention and Correction in Health and Physical Education.

XXI. THE MANUAL AND INDUSTRIAL ARTS 522

I. SOCIAL AND EDUCATIONAL SIGNIFICANCE OF THE INDUSTRIAL ARTS: 267. Manual and Industrial Arts in the Modern Educational Program. 268. Objectives of Industrial Arts Instruction.

II. TESTS IN INDUSTRIAL EDUCATION AND PRACTICAL ARTS: 269. General Uses of Industrial Arts Tests. 270. Measurable Factors. 271. Mechanical Drawing and Visual Arts Tests. 272. Standardized Industrial Arts Tests. 273. Home Mechanics Test. 274. Tests in Household Arts. 275. Aptitude and Intelligence Tests in Industrial Education. 276. Vocational and Trade Tests.

XXII. MEASUREMENT OF GENERAL EDUCATIONAL ACHIEVEMENT 543

I. GENERAL MEASURES OF ACHIEVEMENT: 277. General *vs.* Specific Diagnosis. 278. Advantages and Disadvantages of the Battery-Type Tests.

II. TYPES OF GENERAL ACHIEVEMENT BATTERIES: 279. The New Stanford Achievement Tests. 280. The Unit Scales of Attainment. 281. Metropolitan Achievement Tests. 282. Modern School Achievement Tests. 283. Progressive School Achievement Tests. 284. Sones-Harry High School Achievement Test. 285. Iowa High School Content Examination. 286. Iowa Every-Pupil Tests. 287. American Council Co-operative Tests. 288. Other General Achievement Batteries.

XXIII. USING THE RESULTS OF TESTING 557

I. PUPIL ADJUSTMENT THROUGH MEASUREMENT: 289. Education as Guidance. 290. Disciplinary

CHAPTER PAGE

Uses of Tests. 291. Classification and Placement of Pupils. 292. Sectioning of Pupils. 293. Evaluation of Teaching Method.

II. MEASUREMENT OF PROGRESS AND IMPROVEMENT BY STANDARDIZED TESTS: 294. Standards and Norms as Essentials in Measurement of Progress. 295. Limitations of Standard Tests as Measures of Improvement.

III. MEASUREMENT BY INFORMAL OBJECTIVE TESTS: 296. Informal Objective Tests as Measures of Class Progress. 297. Limitations of the Subjective Bases of Teachers' Marks. 298. Functions of the Teacher's Mark. 299. Objectifying the Marking System.

XXIV. TESTS AND THE HIGH SCHOOL TEACHER . . 572

300. Recognition of the Need for Tests in the Classroom. 301. Meaning of Educational Tests. 302. Teacher-Made Measures of Achievement. 303. Using Informal Objective Examinations in the Classroom. 304. Uses of Standardized Educational Tests. 305. Using Standardized Tests in the Classroom. 306. Summarizing and Interpreting the Results of Testing. 307. Diagnostic Testing and Remedial Teaching. 308. Measurement and the Total Individual. 309. After Testing, What? 310. Conclusion.

APPENDIX 585

A. Publishers and Distributors of Educational Tests Mentioned in This Volume.

B. List of Educational Tests Mentioned in This Volume . 586

C. Glossary 596

INDEX 607

LIST OF TABLES

TABLE NO.		PAGE
1.	Shifting Standards of Expectancy	49
2.	Comparison of Grades Assigned With and Without the Use of Scoring Rules	54
3.	Distribution of Mental Ability of a 9th Grade Class	94
4.	Marks Assigned by Ten Teachers to an Essay Type Examination in Civics, and True-False Examination Over the Same Material	112
5.	Discriminative Power of Test Items	139
6.	Reliability Coefficients and P. E. of Scores for Iowa Silent Reading Tests; Advanced	144
7.	Scores on a History Test Taken by an 11th Grade Class	151
8.	Algebra Test Scores of Thirty-Seven 9th Grade Pupils	152
9.	Scores Given in Table 8 Grouped in Intervals to Form a Frequency Distribution	157
10.	Relation of Range and Size of Class-Interval	158
11.	Distribution of Scores and the Calculation of the Arithmetic Mean	165
12.	Algebra Test Scores Arranged in Descending Order	168
13.	Algebra Scores of a Ninth Grade Class	169
14.	Data Showing Identical Means; Unlike Variability	171
15.	A Frequency Distribution	175
16.	Standard Deviation from Ungrouped Data	179
17.	Standard Deviation from Frequency Distribution	182
18.	Standard Deviation Technique for Assigning Class Grades	184
19.	Pairs of Test Scores	190
20.	Illustration of Computation of XY Products	193
21.	Correlation Table	194
22.	Per Cent of Forecasting Accuracy for Values of *r*	198
23.	Relative Ranks	202
24.	Computing Deciles	204

TABLE NO.		PAGE
25.	Interpretation of Percentiles	205
26.	Differential Mental-Age-Grade Norms	218
27.	Sampling of Scores on Grammar Information Test	219
28.	Percentile Norms Smoothed	220
29.	Distribution of *Iowa Silent Reading Comprehension* Scores by Chronological Ages	225
30.	Age-Equivalents of Composite Scores on *New Stanford Achievement Tests*	227
31.	Pupil Record from *Compass Diagnostic Test* No. 7	229
32.	Grade Norms for *Compass Diagnostic Test* No. 7	229
33.	Age-Equivalents of *Compass Diagnostic Test* No. 7	230
34.	Comparison of Units of Measurement Used in Two Reading Tests	231
35.	G-Scores for *Thorndike-McCall Silent Reading Scales*	233
36.	C-Scores and Reading Ages for *Unit Scales of Attainment*	237
37.	Data for Reading Ages	238
38.	Data for Reading Quotients	241
39.	Data for Accomplishment Quotients	242
40.	Record of Test Scores of Pupil	252
41.	Arrangement of *Kuhlmann-Anderson Tests*	267
42.	Heinis Mental Growth Units	277
43.	Median Intelligence Quotients Necessary for Success	293
44.	Typical Scores for Occupational Groups on Army Alpha	293
45.	Reliability and Validity of Minnesota Mechanical Ability Tests	304
46.	Iowa Elementary Language Test Content	324
47.	Rate Standards for Preparing Timed Dictation Exercises	335
48.	Grade Standards in Writing	339
49.	Classification of Items in Iowa General Information Test in American History	375
50.	Content of Gregory-Owens Mediaeval and Modern History Test	380
51.	Wiedefeld-Walther Geography Test	384
52.	Kinney Scale of Problems in Commercial Arithmetic	396
53.	Content of Breslich Algebra Survey Test	399
54.	Iowa Algebra Aptitude Test	403
55.	Correlation of Algebra Aptitude Tests	404
56.	Lane-Greene Unit Achievement Tests in Plane Geometry	408

TABLE NO.		PAGE
57.	Iowa Plane Geometry Aptitude Test	411
58.	Ruch-Popenoe General Science Test Subject-Matter Emphasis	424
59.	Content of Kwalwasser-Dykema Music Tests	477
60.	Content of Kwalwasser-Ruch Test of Musical Accomplishment	478
61.	Content of Providence Inventory Test in Music	480
62.	Lewerenz Test in Fundamentals of Visual Art	487
63.	Health Education Tests	500
64.	Reliability of a Single Form of Newkirk-Stoddard Home Mechanics Test	533
65.	Norms for Newkirk-Stoddard Home Mechanics Test	533
66.	New Stanford Achievement Test; Advanced Examination	548
67.	Unit Scales of Attainment	549
68.	Illustration of Arrangement of Reading Test Items in Unit Scales of Attainment	550
69.	Reliabilities of Modern School Achievement Tests	551
70.	Data on Progressive School Achievement Tests	552
71.	Data on Sones-Harry High School Achievement Test	552
72.	Data on Iowa High School Content Examination	553
73.	Data on Iowa Every-Pupil Tests	554
74.	Scores from a Ninth Grade Class on Iowa Algebra Aptitude Test	561
75.	Marks Assigned a Sixth Grade Geography Paper	565
76.	Suggested Point Values Corresponding to Letter Grades	570

LIST OF FIGURES

FIGURE		PAGE
1.	Analysis and Diagnosis	38
2.	Illustration of Influence of Sampling on Test Scores	63
3.	The Principle of Sampling	109
4.	Mid-Points and Limits of Class-Intervals	156
5.	Illustrating Principle of Moments of Forces	161
6.	Illustrating the Calculation of the Mean	164
7.	Illustrating the Need for Measures of Variability	172
8.	Comparison of Standard Deviation and Quartile Deviation	177
9.	Percentile Graph	206
10.	Graphic Presentation of the Median, 25th, and 75th Percentiles	222
11.	Comparison of Handwriting Quality with Norms	224
12.	Illustrating Need for Refinement of Age Norms	226
13.	Pupil's Graphic Record	236
14.	Item 125. Meier-Seashore Art Judgment Test	485
15.	Scoring Key for Lewerenz Test in Fundamental Abilities in Visual Art, Test 5	489
16.	A Program to Protect and Conserve Health	495
17.	A Program to Improve Health and Social Efficiency	497
18.	Relation of Health and Physical Education	498

LIST OF PROBLEMS

I. Tabulating Test Scores

PAGE

1. Physics Test Scores of 21 Twelfth Grade Pupils 159
2. Reading Test Scores of 41 Ninth B Grade Pupils 159
3. Language Test Scores of 31 Tenth Grade Pupils 160

II. Finding the Arithmetic Mean

4. Computing the Arithmetic Mean from Frequency Distribution . 166
5. Computing the Arithmetic Mean 167
6. Computing the Arithmetic Mean 167

III. Finding the Median

7. Computing the Mid-Measure 170
8. Finding the Mid-Measure . 170
9. Tabulating, Finding Median, and Arithmetic Mean 171

IV. Expressing the Variability of Test Scores

10. Finding the Quartile Deviation 176
11. Computing the Standard Deviation 187
12. Computing and Assigning T-Scores 187
13. Using Standard Deviation for Assignment of Letter Grades 187

V. Expressing Relationships

14. Computing Correlation of Odd-Even Scores as Basis for Reliability Coefficient 200
15. Computing Correlation of First and Second Scorings of Test as Basis for Objectivity Index 201

VI. Expressing Relative and Absolute Ranks

		PAGE
16.	Assignment of Relative Ranks	203
17.	Computing Percentile Equivalents of Test Scores	207
18.	Preparing a Percentile Graph for Test Scores	207

VII. Using Tables of Norms

19.	Comparing Test Scores with Grade Norms	244
20.	Finding Age Equivalents from Test Scores	244
21.	Finding Quotients from Given Ages	245

The Use and Interpretation of HIGH SCHOOL TESTS

CHAPTER I

INTRODUCTION

THE purpose of this chapter is to introduce the reader to the following points underlying the idea of measurement in education:

a. Measurement in education not a recent movement.
b. In the minds of pioneers many years prior to its general public acceptance.
c. The real pioneers in the measurement movement.
d. The measurement idea as affected by studies of the inaccuracy of teachers' marks.
e. School surveys and the development of tests in school subjects.
f. The stimulation of the measurement movement by the organization of research bureaus.
g. The minimal principles needed in the use and interpretation of tests.

BACKGROUND OF THE GROWTH OF THE MEASUREMENT IDEA

1. **Educational Measurements Not New Idea.** Educational tests and the types of information resulting from their use in the classroom are coming to be almost universally identified with good teaching practice. In many ways teachers have always endeavored to measure the progress of their pupils toward an educational goal and to diagnose revealed defects in instruction. This measurement of progress and the accompanying diagnosis have been in the past largely a matter of observation on the part of the teacher. And yet the recent development of educational tests may be regarded as an extension and improvement of an old practice. However, objective tests are more precise and exact than the ordi-

nary teacher's examination and thus they accomplish the purposes of measurement much more reliably. They permit the setting up of specific and objective goals of achievement which are based upon the actual attainments of children under typical school conditions. They provide answers to such specific questions as "How much should be expected of a given class or pupil under certain school conditions?" or "How much progress should be made in a given subject in a given period of time?"

2. **Real Pioneers in Measurements.** The idea lying back of the standardized objective test is relatively simple. It involves mainly the careful selection and repeated use of certain examination exercises by groups of individuals until the typical reaction of these individuals becomes known. These typical reactions later are designated as norms or standards in terms of which other results can be compared and evaluated. In view of this relatively simple situation it is strange that the idea of measurement has been so long in establishing itself generally in the teaching field. There is evidence to indicate that the idea had been comprehended in part by teachers before it was developed and popularized by modern workers. Professor E. L. Thorndike has called attention [1] to a plan for measuring the achievements of pupils which was used by Rev. George Fisher, headmaster of a school in Greenwich, England, as early as 1864. The instrument used was called a "scale book" in which were assigned values to each degree of proficiency in different school subjects. It was naturally somewhat crude in its organization, but it included the germ of many of the ideas which are incorporated in our present-day educational tests. Fisher's work, like that of most pioneers, produced no lasting results because he lived too far in advance of the thought and educational practice of his day. His work appears to be isolated from the development of educational tests in this country.

The early development of standardized objective tests for the measurement of pupil achievement was contributed to by workers in the fields of both psychology and school administration. The psychologist made measuring instruments for

[1] *Journal of Educational Psychology*, Vol. 4, p. 551.

studying the functions of the human mind. The school administrator was interested more particularly in their relation to problems of the course of study, school organization, methods of instruction, and supervision. In this connection it should be noted that the construction of many of our present-day measuring instruments was not possible until the statistical methods needed in the construction of tests had been carefully worked out. In this phase of test development Sir Francis Galton deserves special mention for his early recognition of the need for more exact statistical tools such as were later developed by Pearson, Spearman and others.

In America the real inventor of the comparative test was Dr. J. M. Rice. After studying in Germany, he returned to America with the notion that it should be possible to measure efficiency and compare results in education as in other fields of science. Until after he demonstrated the fundamental truth of his theory, by giving tests in arithmetic and in spelling in a large number of school systems in this country and revealing some startling differences in achievement, very little progress was made in the development of reliable measuring devices. Various types of tests were used for identical purposes, and without analysis of relative difficulty. How could one city or school system compare its work with another under such circumstances? Rice, by giving the same test to pupils in many different cities, took the first significant step toward procuring for teachers more reliable and dependable measuring instruments.

Stimulated by the work of Dr. Rice, Professor Thorndike began to experiment with tests and scales and to preach the need of such instruments. Stone, Hillegas, Buckingham, Trabue, Courtis, and many others were stimulated by his teaching, with the result that within the decade following Rice's introduction of the measurement idea into America many significant contributions were made to educational measurement. Although the history of testing in this country is a long and honorable one, it is mainly of interest to the historians. For the purpose of building a basis for the better understanding and appreciation of measurement in American education it may be sufficient merely to point out

that in general the productions and findings of many of these early workers were denounced both by the press and the public as foolish and impracticable. During the last twenty years, however, this attitude has quite definitely changed, and an understanding of objective measurement procedures has come to be an essential phase of the technique of scientific education and a necessary part of the teacher's professional equipment.

3. **Factors in Growth of Measurement Idea.** The production of educational tests has been indirectly stimulated by at least three developments:

(1) The numerous important studies of the accuracy of school marks[2] showed the marks to be highly subjective,[3] and hence inaccurate. This accumulation of evidence was conclusive proof not only that such subjective measures as teachers' marks might be expected to vary widely from teacher to teacher but also that the same teacher's marks would vary from time to time and from subject to subject. It was shown that very often success or failure in a given subject depended largely upon the pupil's fortune in being placed in one teacher's section of the class rather than in that of another. This type of demonstration revealed the need for instruments which would yield more accurate measures of achievement.

(2) In the early part of this century a prevailing feeling of dissatisfaction with existing educational conditions coupled with a curiosity and a desire to reveal true conditions on the part of a few educators led to an unusually large number of intensive investigations or surveys of certain of the larger school systems. This survey movement stimulated the construction and use of educational tests, and has in turn been largely influenced by the derivation of more objective devices for measuring the abilities of school children.

(3) The development of educational measurements has been greatly facilitated by the establishment of research organizations. Research bureaus have been organized in many of the larger school systems primarily for the purpose of directing and stimulating the application of educational tests. Similar research bureaus have been established in a large

[2] See Selected References at end of chapter.

[3] See *Glossary* beginning on page 596 for meaning of this and other new terms.

number of state educational institutions, and recognition has been given to the work by the department of public instruction in a number of states. This type of research organization has been very influential in popularizing the use of educational tests. Through these bureaus, school administrators, supervisors, and teachers, have been made acquainted with tests, and frequently have been trained in their use. Service has also been rendered by making tests easily accessible, and by standardizing tests devised by other workers. In many instances the administrators of the various research organizations have been directly responsible for the derivation of many new testing instruments.

4. **Rapid Growth of Interest in Tests.** There are now tests and measuring devices for practically all subjects in the elementary and high school field, as well as in many branches of higher education. Mental tests and all kinds of prognostic, diagnostic, and placement[4] tests are rapidly increasing in number and improving in quality. No one can predict the ultimate range of application and availability. Although the advantages of objective tests and measurements are generally recognized by educators and laymen, many of their shortcomings are as thoroughly realized. Tests are means to an end. Their value depends upon certain characteristics innate in the tests themselves, upon the proper utilization of their results for evaluating achievement, for analysis, diagnosis, and the development of an effective follow-up or remedial program.

The recognition of the value of standard tests for diagnostic purposes has done much to popularize their use and to transfer them from the hands of specialists to the hands of the classroom teacher. When tests were new and still in the experimental stage it was probably desirable for persons other than the classroom teacher to administer them. But now that their value has been demonstrated it is important that teachers master the technique of using tests to locate specific weaknesses of individual pupils. It is also recognized that teachers should be able to select and apply the proper preventive and remedial measures.

[4] See *Glossary*, beginning on page 596 of *Appendix*, for further interpretation of meaning of terms.

5. Purpose of This Book. The purpose of this book is two-fold: (1) to interest the student of secondary education in the possibilities of measurement in the high school, and (2) to stimulate the high-school teacher and supervisor to make more effective use of tests as integral parts of enlightened teaching practice. To accomplish this two-fold purpose the reader is gradually introduced to the meaning and possibilities of measurement through the examination of some of the well-known current classroom practices. Chapters II and III show what tests are and how they provide the basis for the teacher's improvement in his own methods of measuring classroom achievement. The application of modern objective methods of measuring the results of classroom instruction is made in Chapter IV. The uses of standardized tests, the general rules for carrying out a testing program, as well as fundamental principles which are used in the construction of standardized tests are presented in Chapters V, VI, and VII.

The high-school teacher, successful in the use of tests in his teaching, must be able to secure an accurate summary and interpretation of results. The discussion in Chapters VIII and IX is presented for the purpose of developing these skills. The simple statistical procedures such as making a frequency table, calculating medians, arithmetic means, quartiles, etc., will be found especially helpful in interpreting results where fairly large numbers of cases are concerned. The discussion of derived scores in Chapter IX should also be useful to those making an analysis and interpretation of test results. Experience with practical problems of this nature is provided by exercises which are scattered through these two chapters. Those wishing to follow the use of classroom tests to its logical conclusion will wish to study with particular care Chapters X to XXIV inclusive in which are presented the possibilities of measurement in the basic high-school subjects, and the practical methods of using test results for analyzing and correcting many of the learning difficulties of high-school pupils.

The student who is genuinely interested in improving his understanding of many of the points presented in this volume will find much profit in the careful preparation of

the discussion exercises at the end of each chapter. Those who are still more deeply interested and wish to study further into the problems of measurement in education will find the selected references at the close of each chapter of particular value. Because the field of educational measurements is so rich and the essential material is so extensive, it is impossible to compress into the few pages allotted to this book much of that which is considered by many to be fundamental.

Teachers themselves expert in the technique of learning know that passive reading, while yielding information and appreciation, does not develop easy dependable skill in doing the thing described. To provide the student and teacher with an opportunity to secure a more complete mastery of certain of these techniques, a *Work-Book in Educational Measurements* has been prepared as a companion volume for this treatment. In this *Work-Book* the reader may solve actual and practical problems of the type which the classroom teacher and supervisor face. Mastery of the theory of measurement in this textbook and a careful working of the forty-five projects in the *Work-Book* will practically guarantee to the reader an actual concrete experience with the major problems of a dynamic testing program as one phase of the improvement of classroom instruction.

Discussion Exercises

1. What specific evidence do you find that the idea of measurement in education is not an entirely new idea but has been in the minds of teachers for many years?
2. In what definite ways have the various studies of inaccuracy in teachers' marks affected the attitude of teachers toward objective measurements?
3. How far, in your opinion, is the classroom teacher responsible for the understanding and use of educational tests?
4. What three important educational developments of recent years have indirectly stimulated the growth of interest in educational measurements?
5. Who were the real pioneers in the development of objective tests? What seems to be their characteristic influence on the measurement movement?
6. What are the reasons why some of the early efforts at objective measurement, such as that represented by Fisher's *Scale Book*

of 1864, failed to hold the attention of later groups of teachers?

Selected References

Ayres, L. P., *History and Present Status of Educational Measurements.* 17th Yearbook of the National Society for the Scientific Study of Education, Part 2. Bloomington, Illinois: Public School Publishing Company.

Thorndike, E. L., "Handwriting Scales" in *Teachers College Record,* Vol. 2, No. 2. March, 1910.

CHAPTER II

THE MEANING OF MEASUREMENT IN EDUCATION

THE following points involved in securing a clear notion of the meaning of measurement in education are discussed in this chapter:

a. What measurement in education is.
b. What tests are and do.
c. Qualities which distinguish standard tests from informal examinations.
d. Superiority of standard educational tests over other types.
e. Distinguishing features of tests and scales.
f. Differences between rate tests and power tests.
g. Types of mental examinations.
h. Distinguishing features of general survey, analytical and diagnostic tests.

I. WHAT TESTS ARE

6. Measurement in Education. In general, in any field of human endeavor, the most efficient results are attained when the worker has definite goals toward which to work, and dependable instruments for determining progress. A definite aim enables the worker to direct his efforts more efficiently toward the particular task to be accomplished. By the proper use of instruments for measuring results it is possible for the worker to know what he has accomplished. Thus a reliable and analytical silent reading test will give to the teacher a measure of his relative success in developing the important silent reading skills in his classes. Accurate measuring instruments also aid in discovering when emphasis has been misplaced. For example, a pupil who in his elementary school work has been given an unusual emphasis on oral reading may satisfactorily pronounce words appear-

ing on the printed page, but may be lacking in ability to get meaning from these same words.

Measuring instruments also make it possible for the worker to resort to experimental methods and thus learn definitely whether materials and methods are effective. This is as true in the field of teaching as in other fields. Without specific aims the teacher cannot plan his work effectively. He cannot know, except in an indefinite way, what he is to do. A teacher without specific aims is like a person who starts out to walk for a certain period without any notion of which direction he is to take or how far away his destination may be. If, on the other hand, the goals of instruction are clear-cut and accurate, means of determining progress are provided, and the probability of a timely arrival at the goal is greatly increased.

7. **Characteristics of Educational Tests.** Modern educational tests differ in several respects from the typical examination of the discussion type such as teachers commonly construct. *First,* the questions or exercises which are used in educational tests are often much more carefully selected to coincide with the purpose for which the test is designed than is true of the ordinary informal examination. For example, an objective or a standard test having for its purpose the measurement of ability to locate the states of the United States contains only such exercises as relate specifically to that purpose. The traditional informal examination frequently contains exercises from many fields. *Second,* the exercises in a carefully made educational test are commonly arranged in accordance with certain principles of test construction so as to form an accurate measuring instrument. These principles of test construction, being quite technical, are discussed in a later chapter in this book. For the present it is enough for the reader to note that there are important principles of arrangement of items within a test which should be considered.

In the *third* place, the test or scale is standardized, i.e., the test is given to a large number of children of varying age and school classification. From these results the norms or standards are obtained. The interpretation of norms and standards and the meaning of certain scores derived from

these values is discussed in Chapter IX of this book. A *fourth* point of great importance, but only recently recognized and applied in connection with test development, is the fact that test scores yielded by narrow-function tests are more readily translatable into specific remedial procedures where needed. This is nothing more than saying that a really valuable test, in addition to giving a cross-section of the situation, must diagnose, and that this diagnosis must be so specific that the results may be readily interpreted in terms of the specific kind and approximate amount of remedial drill needed. A moment's thought will show the reader that a busy teacher can hardly be expected to construct tests possessing the values of standard tests. Neither should the teacher be forced to depend upon his general powers of observation for this specific type of information about his students. A clear appreciation of the possibilities of tests as integral parts of superior teaching practice will reveal that the teacher's most important task in this regard is the selection of the tests best suited to his needs, just as it has long been his duty to select the textbooks best suited to his purposes. In a similar way the teacher's relation to preventive and remedial materials and measures should be one of selection and application rather than one of construction.

8. What Tests Do. Probably every alert teacher has at some time earnestly desired to know whether instruction in certain school subjects was particularly superior or inferior, effective or ineffective. Many teachers have taken steps to answer this highly important question through the use of one kind of measuring instrument or another, yet relatively few classroom teachers are utilizing to the fullest extent one of the most valuable instruments at their command. Many supervisors and administrators are frequently obliged to admit that they did not have adequate data on which to base their decision as to the efficiency of a certain method of instruction but had to be guided by their own personal opinion. With the development of more valid and reliable measuring devices as a part of the scientific technique of education this will become more and more unnecessary.

Standardized tests are not panaceas for all educational inadequacies but unquestionably the scores they afford are

useful in the evaluation of instruction. Test scores are objective, and may be given meaning by the process of standardization. For example, a quality score of 46 assigned to a certain handwriting sample is meaningless until it is understood that 46 points is standard quality for a fourth-grade child. If such a sample were scored as a second-grade product it might be assigned a superior mark. If graded by an eighth-grade teacher it might readily be given an inferior mark. Thus, by the use of standardized test scores, specific goals of achievement may be set up and progress may be measured. Test norms and standards themselves provide the basis for the more objective grade placement of pupils. Analytical and diagnostic tests make possible the discovery of pupils needing special corrective instruction. Weak spots in the course of study may be pointed out, methods of instruction may be evaluated, and general school efficiency may be improved through the critical use and interpretation of educational tests.

Experience in the use of tests in many school systems leads to the conclusion that quite often there is an intangible psychological effect resulting from the administration of a series of tests in a school. The experience of the children while taking the test, and the feeling on the part of the teacher that his work is being carefully checked are both motivating forces making for better and more effective teaching and learning situations. This is, however, only a by-product of the use of the test and is no reason for allowing the work to stop short of a really constructive supervisory program.

9. **What Tests Do Not Do.** Standard tests are incapable in and of themselves of directly improving instruction in any subject. They merely reveal the situation. In a sense, a test might be thought of as an educational barometer. It merely reveals the educational-atmospheric pressure but does not do anything about it. Perhaps the parallel is closer than at first appears. Just as low barometric readings indicate low atmospheric pressure and forecast changing weather conditions or storms, low achievement test scores may presage an unsatisfactory educational situation. As high or rising barometric records indicate fair weather, so high re-

sults on educational tests indicate a satisfactory instructional situation.

The chief service of tests is their power to reveal the strengths and weaknesses of individual pupils or of the class as a whole. Their use must be followed by the next logical step,—the development of a constructive supervisory program. It is not enough that weaknesses be revealed. They must be corrected by the use of properly constructed remedial exercises based upon an adequate psychology of learning.

II. Kinds of Tests

10. **Essay-Type Tests.** The two kinds of educational tests in common usage are (1) objective and (2) essay-type tests. While both of these classes of tests are discussed quite completely in later chapters it is worth while to describe each briefly at this point. Objective tests are so constructed that they can be scored with a minimum of guessing or subjective judgment. The student's responses to the exercises are brief and specific. In the traditional or essay-type test a limited number of questions (usually five to ten) are stated by the teacher as the basis for discussion by the student. Often the questions are stated hurriedly, and frequently they are selected without special attention to the importance of the subject matter. The result is a vague type of discussion of exercises which themselves represent only a brief and careless sampling of the subject matter for which the students should be responsible.

The scoring of the essay-type examination represents a real problem. Many factors appear which influence the teacher's judgment but have little to do with the actual evaluation of the student's knowledge of the subject. Some of these factors are English, including spelling, sentence structure, paragraphing, composition; mechanical features of the examination, such as neatness, legibility, form; the teacher's attitude toward the individual pupil, or the pupil's attitude toward the teacher. These factors operate in unknown combinations to influence the final mark. This means that the mark on the test is an entirely inadequate

expression of any one factor, and hence is probably an unreliable measure of the entire field covered by the test. Thus the essay-type test, constructed and used under these conditions, can furnish only a very rough measure of achievement. There are, however, certain conditions under which it does function quite effectively. This general criticism of the essay-type test does not necessarily condemn it for all measurement purposes in the classroom. Many more of the limitations and advantages of this type of test are presented in Chapter III.

11. **Objective Tests.** Many of the conflicting factors which appear to invalidate measurement based on the essay-type question are quite largely eliminated in properly constructed objective test exercises. Objective exercises are marked by two important features; namely, (1) brevity of pupil response, and (2) complete absence of personal judgment in scoring the exercises. These two features of the objective exercises make it equally suitable for use in the teacher-made informal examination, or in the more carefully constructed and refined standardized test.

In the objective-test situation the pupil indicates his response by the very briefest and simplest of physical reactions. Usually the record of the student's reaction consists in the mere underlining or encircling of a word or a phrase. Because of this brevity of response many more reactions can be given by the pupil in a unit of time with the very important result that a much more complete coverage or sampling of the subject matter is afforded. The use of the objective form of test item enables different individuals to score the test papers and secure identical results. The fact that a single acceptable answer to a test item is indicated in the answer key practically guarantees consistency in the scoring of such types of test items. A test exercise which is perfectly objective may be scored repeatedly by the same or by different individuals without significant variation in the results. Such accuracy in grading of test items can be obtained only when the exercises are constructed in accordance with certain rather well-known specifications. The use of the objective forms of test exercises eliminates many of the undesirable features of the traditional examination, such

as (1) the variation in the length of the pupil's response, (2) the limitation on the sampling of the subject matter, (3) the number of questions to which the pupil may respond due to the interference of the mechanics of writing, and (4) the general difficulty in evaluating the accuracy and worth of the pupil's statement.

A complete discussion of the advantages and disadvantages of objective tests, and of the specifications lying back of their preparation, accompanied by illustrations of different types, is given in Chapter IV.

12. **Tests and Scales.** Educational measuring instruments are roughly divided into tests and scales. This distinction is of some value, but at times it results in confusion because certain types of objective tests resemble scales or contain certain features of scales as an essential part of their construction. In general terms, a *test* is a measuring instrument designed for the evaluation of any knowledge, quality, or ability. It may measure degree or amount of achievement, mental abilities, aptitudes, or even more indefinite qualities such as character traits. It may be made up of items of similar difficulty, or it may be composed of items of uniformly increasing difficulty or value. In the former case it is a *rate* test; in the latter case it is a *power* test. The process of determining the difficulty or value of test items used in the more refined instruments is called *scaling*. The use of this term in both the noun and the verb senses may account for much of the confusion concerning tests and scales.

The term *scale* is used to designate a series of objective forms of exercises or definite samples of products of different quality which by means of a rather technical statistical procedure have been arranged in a definite order or position, usually in ascending order of difficulty or merit. In a scale each exercise or group of exercises of a specific value or difficulty (or quality) represents as much greater value and higher merit than the exercise or group just below on the scale as the next following exercise or group is more difficult or higher in quality than the exercise or group just below it on the scale. The term *scaled* as applied to test exercises means that the exercises have been submitted to groups of students for their responses, and have been arranged in order

of difficulty as indicated by these reactions. In such a scaling process, failure of pupils classified in definite age or grade groups to respond correctly to the test item is taken as evidence of its difficulty.

An educational measuring instrument is a *scale* to the extent that it ranks accomplishment directly in terms of systematic levels, grades, or ages. An instrument which is made up of scaled items (items of systematically increasing difficulty or merit) and which expresses its results in terms of the number of such items responded to correctly is still a test. Such a test is made quite often by the process of selecting items of known value or difficulty from a scale. For example, a spelling test made up of words of gradually increasing difficulty could be made from the *Simmons-Bixler High School Spelling Scale* (see sample on page 19) by selecting as test words those words which a uniformly decreasing percentage of pupils spell correctly. This test might be treated as a scale if the scores on it were expressed in terms of the value of the most difficult word spelled correctly. If the results were expressed in terms of the total number of these words spelled correctly it would be considered a test.

Most of the more recent tests are really hybrids resulting from combination of the power idea and the rate idea in testing. That is, they are made up of scaled items arranged in ascending order of difficulty, but the resulting scores representing accomplishment are expressed in terms of the number of items responded to correctly in the specified working time regardless of the difficulty.

The scaled type of test is nicely illustrated in a sampling of exercises taken from the *Analytical Scales of Attainment in Literature*, Division 4.[1] These exercises represent the first five and the last five items in the part of the test designed to be used in the ninth grade. These exercises are reproduced on page 17.

[1] Published by Educational Test Bureau, Minneapolis, and Philadelphia. Developed under the authorship and editorial direction of M. J. Van Wagenen, University of Minnesota.

ANALYTICAL SCALES OF ATTAINMENT IN LITERATURE

DIVISION 4, NINTH GRADE

INFORMATION ABOUT LITERATURE

Directions: Read these two sentences carefully.

A. In the story of "Joseph and His Brothers," Joseph's brothers treated him very **1. kindly 2. cruelly 3. justly 4. nobly 5. generously**.. A. ......2......

B. Robinson Crusoe was **1. lost on the ocean 2. left by his sailors 3. lost in a desert 4. shipwrecked on an island 5. seized by bandits**.. B. ....

You see that there are five possible answers in each sentence. Only one answer is right. In the first sentence the right answer is **cruelly**, so a line is drawn under **cruelly**, and the number in front of it, 2, is put at the end of the line.

Now look at the second sentence above and listen to the next directions.

In each of the following sentences you are to find the right answer, draw a line under it and then put the number that is in front of it at the end of the line, just as in the samples above.

---

1. A story about people left on an uninhabited island is **1. Deerslayer 2. Swiss Family Robinson 3. Men of Iron 4. Call of the Wild 5. Westward Ho**.............................. 1. ....

2. The "Just So Stories" are mostly about **1. painted dolls 2. children 3. witches 4. fairies 5. animals**................ 2. ....

3. Folk tales of colored people are found in **1. Tanglewood Tales 2. Odyssey 3. Uncle Remus 4. Arabian Nights 5. Arthur and His Knights**.............................. 3. ....

4. A scene where an emperor is stabbed to death occurs in **1. Julius Caesar 2. Ivanhoe 3. Men of Iron 4. The Spy 5. Quentin Durward**.............................. 4. ....

5. The "Odyssey" tells of the wanderings of **1. Siegfried 2. Odin 3. The Black Knight. 4. Ulysses 5. Hiawatha**...... 5. ....

36. "Maud Muller" tells about **1. the beauties of nature 2. life in the country 3. an heroic action 4. an unsatisfied longing 5. joys of childhood**.............................. 36. ....

37. The events of "The House of Seven Gables," take place in **1. Virginia 2. Louisiana 3. Michigan 4. Massachusetts 5. Maine**.............................. 37. ....

38. Rebecca is a character in **1. House of Seven Gables 2. Lady of the Lake 3. Ivanhoe 4. Ramona 5. The Spy**........... 38. ....

39. Hawkeye is an important character in **1. Treasure Island 2. Last of the Mohicans 3. By Pike and Dyke 4. The Texan Triumph 5. By England's Aid**.............................. 39. ....

40. In which one of these books do birds play a large part? **1. Master Skylark 2. Westward Ho 3. Call of the Wild 4. Ben Hur 5. Wonderful Adventures of Nils**............... 40. ....

Number right..........

The Simmons-Bixler Standard High School Spelling Scale[2]

The accompanying sample page represents the beginning of Part II of the Simmons-Bixler Spelling Scale and shows the arrangement of the words in alphabetical order with the percentages of spelling accuracy for each word for grades nine to twelve. The unmarked words are those found in the first five thousand of the *Horn Basic Writing Vocabulary*.[3] Words preceded by an asterisk are found in the second five thousand. Words preceded by a dagger have a position beyond the first ten thousand words.

The term *scale* is also applied to certain devices designed to measure the quality or merit of certain types of school products. Good examples of scales of this type are the *Ayres Handwriting Scale* and the *Willing Composition Scale*. In these cases the scales are not sources of material for the construction of tests, but are in themselves instruments for assisting the teacher in the assignment of quality grades to the pupil's production. The scales themselves are not taken into the classroom and are not used as the basis for pupil reactions. They represent more or less objective standards of quality or merit. Scales of this type consist in a series of specimens of the particular product under consideration arranged in ascending order of merit from very poor (or zero quality) to very high quality. The student's product to be evaluated is secured under standardized conditions. The quality of his product is then expressed in terms of the value of the specimen on the scale which most nearly matches his own product. Obviously, the use of such quality scales introduces a certain amount of subjectivity into the measurement, since the teacher's judgment is necessarily involved in assigning the quality rating. However, experience in the use of such scales indicates that a short period of training in their use reduces the unrelia-

[2] Simmons, E. P., and Bixler, H. H. *Standard High School Spelling Scale.* Atlanta, Ga.: Turner E. Smith and Co., 1925. Reproduced with the special permission of the publishers.

[3] Horn, Ernest, *A Basic Writing Vocabulary.* University of Iowa Monographs in Education, First Series, No. 4. Iowa City: University of Iowa, April 1, 1926.

SIMMONS-BIXLER SCALE ARRANGED ALPHABETICALLY

	IX	X	XI	XII
A				
*abandon	71	76	80	84
*abandoned	63	67	71	74
ability	86	90	94	98
absence	80	82	84	86
absent	89	91	92	93
absolutely	85	86	87	88
absurd	56	65	74	83
*abundant	69	79	88	97
*academy	59	67	74	81
*accent	73	81	89	97
accept	89	92	94	96
acceptable	75	80	85	89
acceptance	80	85	89	93
accepted	86	88	90	91
accepting	87	90	92	94
*access	80	84	88	91
accident	89	92	94	96
accidents	85	89	92	95
accommodate	54	65	75	85
accommodation	36	39	42	45
accompanied	64	70	76	82
accompany	88	89	90	91
accompanying	59	63	66	69
accomplish	89	92	94	96
accomplished	85	88	91	93
accordance	87	91	95	98
*accorded	75	83	91	98
accordingly	87	91	95	98
*accounted	75	79	83	87
accounts	87	92	96	100
*accredited	64	72	80	88
accrued	26	29	33	36
*accumulate	59	67	75	83
accurate	83	85	87	90
*accurately	80	83	86	88
*accuse	78	85	91	97
accused	76	83	90	97
accustomed	60	67	74	81
ache	87	91	95	98
*achievement	68	71	74	77
*acid	85	90	95	100
acknowledge	85	87	89	91
acknowledged	72	79	85	91
acknowledging	67	75	83	90
acknowledgment	44	46	48	49
*acquaint	75	81	86	91
acquaintance	67	70	73	75
acquainted	72	79	85	91
acquire	82	84	86	88
acquired	67	69	71	72
acres	89	93	97	100
across	88	89	90	91
activities	81	85	89	93
activity	87	91	95	98
*actor	85	89	93	96
acts	89	92	94	96
actual	89	93	97	100
actually	79	86	92	98
†actuates	41	48	54	60
†acutely	66	68	69	70
†adage	22	33	44	55
†adaptability	69	69	69	69
adapted	81	85	89	93
†adapting	77	81	85	88
additional	82	88	93	98
adequate	45	59	72	85
*adieu	46	50	55	59
†adjoin	79	85	91	97
*adjoining	77	83	89	95
*adjourn	62	68	74	80
adjourned	58	68	77	86
*adjournment	60	67	74	80
*adjustable	84	89	93	97
adjusting	88	90	92	94

bility of measurement resulting from personal bias to a point where it is not seriously significant.

13. **Standardized vs. Informal Objective Tests.** Educational tests of the objective type are further classified as (1) standardized tests and (2) non-standardized or informal tests and scales. A test is *standardized* when it is composed of exercises which have been selected in the light of current teaching emphasis and curricular content, when these exercises have been statistically evaluated as to innate difficulty, and when the test itself is accompanied by norms or standards permitting the interpreting of the results of pupil reactions to the test in terms of levels of accomplishment. Standard tests are of value in making comparisons of the achievement of students with general standards and in comparing groups in different schools with each other and with groups from other school systems. Teacher-made or informal objective tests are similar to standardized tests except for two rather important points. Usually the items in the teacher-made tests more closely parallel the local course of study and the local instructional emphasis than is possible in the case of the standardized test. Generally no norms or standards are available for the interpretation of the informal objective examination. However, such tentative norms may be developed by recording and preserving the results each time the test is given in class. Teacher-made objective tests are extremely useful in the general measurement of classroom achievement. The teacher will find that an adequate mastery of the methods of constructing, using and interpreting the results of informal objective tests will constitute one of his greatest professional accomplishments.

14. **Mental Tests.** General mental ability is commonly thought of as the power of the individual to learn. It is usually measured in terms of the extent to which he has acquired a number of specific and largely unrelated abilities. Most intelligence tests consist of a battery of different tests measuring many different fields. The total score on the battery is taken as an index of general intelligence. In a sense, general mental ability may be thought of as a cable composed of many strands and fibers of varying size and quality, each representing some particular phase of ability.

The test of general intelligence is merely a device for taking a cross section of this cable. To the extent that the particular measuring instrument has in it the elements which reveal the large and important strands of the cable the test is valid, and more exactly reveals the true mental status of the individual measured.

The accurate and objective measurement of general mental ability is so important, and the available measuring instruments for use in this field are so numerous and valuable that a complete chapter must be given to their consideration. The problems of the measurement of general mental ability are discussed in detail in Chapter XI.

15. **Aptitude and Prognostic Tests.** If the general concept of the measurability of mental ability is accepted then it is possible to develop instruments for the measurement of specific types of intelligence. Devices such as the *Stenquist Mechanical Aptitude Tests, O'Rourke Mechanical Aptitude Tests, Detroit Mechanical Aptitude Examination, Seashore Musical Talent Tests, Meier-Seashore Art Judgment Tests, Iowa Algebra Aptitude Tests, Iowa Plane Geometry Aptitude Test, Orleans Algebra Prognosis Test, Orleans Geometry Prognosis Test, Iowa Placement Examinations, Aptitude Series* in the fields of Chemistry, English, Foreign Language, Mathematics, and Physics are supposed to measure the individual's capabilities to learn in the definite field. Such instruments may be called specific intelligence tests. Tests of this kind have also been called prognostic tests.

The accompanying reproduction of a few of the exercises comprising the four parts of the *Iowa Algebra Aptitude Test* may serve to give a more definite idea of the content of such tests. It will be noted that, in general, aptitude tests are designed to measure specific underlying abilities and not achievement in the selected field, although a number of recently developed aptitude tests depend for their prognosis upon the speed and accuracy with which certain processes and skills typical of the special subject can be mastered under controlled study conditions. In the particular test chosen here for illustration the pupil's scores depend upon his control over the selected basic abilities and not at all upon his mastery of algebraic concepts as such.

Specimen Exercises from Iowa Algebra Aptitude Test[4]

GENERAL DIRECTIONS: This test consists of four parts, each with a separate time limit. You will be told when to start and when to stop work on each part. Start and stop promptly when told to do so, in order that you may make as good a score as possible on each part of the work. If you come to an item which you are unable to do, omit it and go on to the next ones.

TEST RECORD

		Time in minutes	Perfect score	Pupil's score
1	Arithmetic	12	30	
2	Abstract Computation	8	25	
3	Numerical Series	12	42	
4	Dependence and Variation	3	10	
	Totals	35	107	

Part 1. ARITHMETIC

Time allowance — 12 minutes

Directions: In this arithmetic test, you are to work each example as directed.

1. Add
2
9
8
4
3
5
2

2. Subtract
978
249

3. Multiply
934
29

4. Divide
8/ 424

5. Add: (All 5 must be correct)
14 + 8 =
15 + 6 =
27 + 9 =
32 + 5 =
38 + 7 =

27. Multiply and point off
20.73
40.3

28. Solve
¾ × 15 =

29. Copy and add
.003 + 4.2 + 7.65 + .17

Copy Example 29 here:

30. Goods marked at 125% of cost sell at $7.50. What is the cost?

[4] Reproduced with permission of the publishers, Bureau of Educational Research and Service, University of Iowa, Iowa City. Copyright, 1931, by H. A. Greene and A. H. Piper.

Part 2. ABSTRACT COMPUTATION

Time allowance — 8 minutes

Directions : The answers to these problems are obtained by adding, subtracting, multiplying, dividing, or by making combinations of these processes. You are to answer the question asked in each problem.

Sample : How many pencils would I have if I bought k pencils, and some one gave me m more pencils ? .... $k + m$

The correct answer is found by adding the k and the m. The answer is $k + m$. In this work be careful to use the correct signs. Begin to work when the signal is given.

---

1. Monday I read a pages of a book, Tuesday I read b pages, Wednesday I read c pages. How many pages did I read altogether ? .... 1.
2. Ray Smith wants to buy a gun which costs x dollars. He has y dollars saved up, but that is not enough. How many more dollars does he need ? .... 2.
3. What is the volume of a room which is r feet wide, s feet long and t feet high ? .... 3.

24. A grain elevator man has 4 bins each capable of holding a bushels of grain. These bins are empty. He is expecting to receive b loads of grain, each load containing c bushels of grain. His present bins will not hold this large amount of grain. Indicate the capacity of the new bin he will need to build in order to have sufficient room for the grain ....24.
25. A newspaper charged $8.00 for the first insertion of a quarter page advertisement, and $4.00 for each following insertion. Express the cost of n insertions ....25.

Part 3. NUMERICAL SERIES

Time allowance — 12 minutes

Directions : Each of the following number series is made up according to some rule. Multiplication, division, addition, and subtraction, and various combinations of these processes are employed in forming the different series. Discover the rule for each example, and write the next *two* terms on the two blank lines.

Sample : 1 2 3 4 5 6 ....*7*.. ..*8*..

The next two terms were 7 and 8, as we were counting by ones in this problem, thus 7 and 8 were put on the two blank lines. Work the following examples in a similar manner.

1.	2	4	6	8	10	
2.	9	8	7	6	5	
3.	1	1	5	5	9	9
41.	1	2	4	12	36	
42.	$\frac{5}{108}$	$\frac{5}{36}$	$\frac{5}{12}$	$\frac{5}{4}$	$3\frac{3}{4}$	

Part 4. DEPENDENCE AND VARIATION

Time allowance — 3 minutes

Directions : Each of the following ten questions can be correctly answered by one of these four phrases. Indicate your answer by using the letter which is in front of the answer which you think is correct.

a. remain the same
b. become larger
c. become smaller
d. cannot tell

Sample : X — 7 = 10 If the 7 were made larger, what change would need to be made in the value of *X* so that the answer 10 would still be correct? ....*b*..

The correct answer would be *b*, meaning that *X* would need to become larger if 7 became larger, and the answer remained a 10. Answer the following questions in a similar manner.

1.	$X = y$	If *y* were made larger, what change would be made in the value of *X* ?	1.
2.	$X = y + m$	If *y* were made smaller, what change would be made in the value of *X* ?	2.
3.	$X = y + \frac{m}{s}$	If *m* were made smaller, what change would be made in the value of *X* ?	3.
9.	$X = y + \frac{m}{s}$	If *s* were made smaller, what change would be made in the value of *X* ?	9.
10.	$X = y + m$	If *m* were made larger, what change would need to be made in the value of *y* so that *X* would not change in value ?	....10.

16. Tests of Educational Achievement. Tests of achievement, or measures of what the pupil has learned as a result of his school or other type of educational experience, are usually classified according to their form or function. In the matter of form most educational achievement tests may be classified as rate or power tests, and quality or product scales. According to function, educational achievement tests may be classified as survey, analytical, or diagnostic tests.

Rate Tests. Rate tests usually consist of exercises approximately equal in difficulty. Ordinarily no pupil is able to finish the test in the working time allowed. The number of exercises finished correctly in the specified time limit is taken as the pupil's rate of work. Thus, rate tests are measures at a given level of difficulty of the speed and accuracy with which a pupil is able to respond to certain standardized

activities of a uniform nature. Usually the exercises are relatively easy and there is little or no question about the pupil's ability to do them. Such a type of test implies that our educational objectives are to train children to extend their power over a larger and larger area of content of constant difficulty. It is doubtful if such an objective is in keeping with the major aims in most subjects. At any rate, there is now a distinct tendency away from the development of tests which are definitely rate or definitely power in their basic characteristics. Most of the better modern tests attempt to combine the best elements of both ideas of measurement into a hybrid form.

Power Tests. Power tests, or the so-called scaled tests (see page 15 of this book for a discussion of this point) measure a pupil's ability to do more and more difficult exercises within a given field of subject matter. In such tests achievement is expressed in terms of the difficulty of the exercise or activity which the pupil is just able to perform. If *rate tests* may be compared to running a race in which a series of hurdles of uniform height are to be cleared, *power tests* may be compared to a race in which the hurdles are very low at the start but become gradually higher and higher as the race goes on until no one is able to clear them. In the first case, the achievement (rate score) would be expressed in terms of the number of hurdles cleared in a specified time. In the second case, the score (ability) would be expressed in terms of the highest hurdle which the individual was just able to clear.

A power test consists in a series of exercises within a given field of subject matter arranged in ascending order of difficulty, and is intended to measure a pupil's power to do more and more difficult tasks of the type included in the test. Usually no measure of the pupil's rate of work is secured because the time allowed in such a test is more than enough for all subjects to complete as many of the exercises as they are able to do. Sometimes, however, rate and quality are combined by taking as the pupil's score the number of exercises of increasing difficulty (or value) done correctly in the specified time. Theoretically a pupil's score on a power test is expressed as the degree of difficulty of the most difficult

exercises he is able to do with a specified degree of accuracy. Such a score is laborious to compute from a pupil's performance, and for this reason the number of exercises done correctly is generally taken as his score.

The characteristics of power tests are illustrated in the accompanying reproduction of the Problems Test from the *Hotz First Year Algebra Scales*, Series A. This is one of the few remaining tests of this type.

17. Quality and Product Scales. *Quality Scales.* Achievement may also be expressed in terms of ratings on quality scales. In such subjects as handwriting, composition, etc., the measuring instruments must obviously be of a different type. In these subjects the pupil produces a performance which cannot be considered as either right or wrong or at least as partially right. The problem becomes one of describing its degree of quality. For this purpose "quality scales" have been constructed which consist of a series of specimen performances. These exhibit varying degrees of quality from the lowest quality to the highest and are arranged systematically in order of increasing quality. Usually the quality of each is described numerically. A quality scale of this type is used by matching the performance to be described with the specimen of the scale which most nearly resembles it in quality. The following quality scales have been widely used: *Willing Scale for Measuring Written Composition*; *Ayres Handwriting Scale, Gettysburg Edition*; *Hillegas Scale for Measuring Written Composition*; the *Thorndike Handwriting Scale*; and *Knight-McClure Arithmetic Neatness Scales.*[5]

Product Scales. In product scales all exercises of a given difficulty or value are grouped together and are properly located along a scale of difficulty. On this account product scales represent excellent sources for rate or power tests. The *Ayres Spelling Scales*, the *Iowa Spelling Scales*, and the *Simmons-Bixler High School Spelling Scales* are examples of this type of material.

18. Survey vs. Diagnostic Tests. *Survey Tests.* Most tests devised for school use aim to measure what the pupil

[5] A list of publishers of tests and scales mentioned in this book is given in the *Appendix.*

SPECIMEN OF HOLTZ ALGEBRA SCALE[6]

SERIES A

FIRST YEAR ALGEBRA SCALES

BY HENRY G. HOTZ

PROBLEMS

Do not work out the answer to the problem — merely *indicate* the answer or *state* the equation in each case.

1. If one coat costs x dollars, how much will 3 coats cost?

   Answer............................

2. A man is m years old; how old was he r years ago?

   Answer............................

4. A gold watch is worth ten times as much as a silver watch, and both together are worth $132. How much is each worth?

   Equation..................................

5. The distance from Chicago to New York by rail is 980 miles. If a train runs v miles an hour, what is the time required for the run?

   Answer............................

7. The total number of circus tickets sold was 836. The number of tickets sold to adults was 136 less than twice the number of children's tickets. How many were sold of each?

   Equation...

8. A rectangular box is d inches deep, w inches wide, and contains r cubic inches. What is its length?

   Answer............................

9. The area of a square is equal to that of a rectangle. The base of the rectangle is 12 feet longer and its altitude 4 feet shorter than the side of the square. Find the dimensions of both figures.

   Equation..................................

12. A train leaves a station and travels at the rate of 40 miles an hour. Two hours later a second train leaves the same station and travels in the same direction at the rate of 55 miles an hour. Where will the second train pass the first?

   Equation..................................

[6] Reproduced with the permission of the publishers, Bureau of Publications, Teachers College, Columbia University, New York, 1920.

has learned. They are accordingly called achievement tests. Many achievements appear to be specific. In such fields as the operations of arithmetic a large number of skills which are relatively independent of one another are recognized. For example, a pupil may attain a very high degree of achievement in the addition of integers without attaining anything like an equivalent degree of achievement in the addition of fractions.

This suggests the possibility of at least two types of achievement tests: first, those which measure certain general achievement; and second, those which yield independent measures of a number of specific achievements within a given field. As illustrations of this first type almost any one of the several hundred achievement tests may be mentioned. Such tests as the *Compass Survey Tests in Arithmetic*, the *Woody-McCall Mixed Fundamentals in Arithmetic*, the *Wilson Language Error Test* have had extensive use in the advanced elementary school grades. Among the more popular batteries of achievement tests for use in the junior high school are the *New Stanford Achievement Tests*,[7] and the *Unit Scales of Attainment*. Other achievement tests such as the *Briggs Form Test in Language*, *Kirby Grammar*, *Van Wagenen's Reading Scales in Literature and in History*, *Hotz Algebra Scales*, *Schorling-Sanford Geometry Tests*, *Powers Chemistry Tests*, *Ruch-Cossmann Biology Tests*, and the *Ruch-Popenoe General Science Tests* have also had extensive use in the high school field.[8] The *Sones-Harry Tests*, a battery designed to afford a general survey of high school achievement, are used quite widely in these grades.

The characteristics of survey tests designed for use in a single subject matter field are shown in the accompanying exercises sampled from the *Ruch-Popenoe General Science Test*.[9]

[7] Kelly, Ruch, and Terman, *New Stanford Achievement Tests*, World Book Company, 1929.

[8] See pages 585-586 for list of publishers of these tests.

[9] Reproduced with the permission of the publishers, World Book Company, Yonkers, New York.

RUCH-POPENOE GENERAL SCIENCE TEST

PART I

DIRECTIONS. Draw a line under the one word or phrase that makes the truest answer to the question, as shown in the sample.

SAMPLE. An example of a liquid is

iron glass air water lead zinc oxygen

Begin here.

1 Pneumonia is a disease of the
heart liver lungs brain stomach muscles kidneys 1

2 Volcanoes are most likely to be found in
deserts valleys coastal plains mountains deltas interiors islands 2

3 Distance above sea level is called
longitude rotation altitude latitude declination revolution inclination 3

49 A metal which can be drawn into fine threads is said to be
elastic ductile flexible malleable amorphous magnetic viscous 49

50 The attraction between molecules of a body is called
capillarity adhesion magnetism inertia buoyancy cohesion convection 50

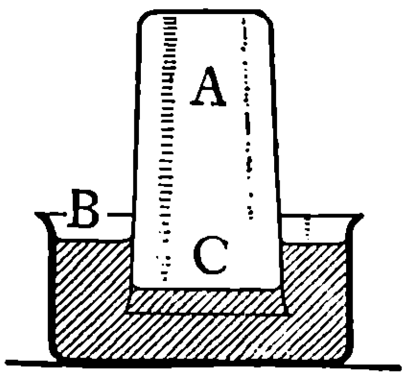
*Score in Part I*____________

PART II

DIRECTIONS. On this page and the following pages there are 20 diagrams and several statements about each diagram. Part of each statement is left blank. You are to fill each blank so as to make a true statement. Make the statements as complete and definite as possible. In most cases you will need to write but ONE WORD in each blank. You will be given 25 minutes to do this part of the test.

Begin here.

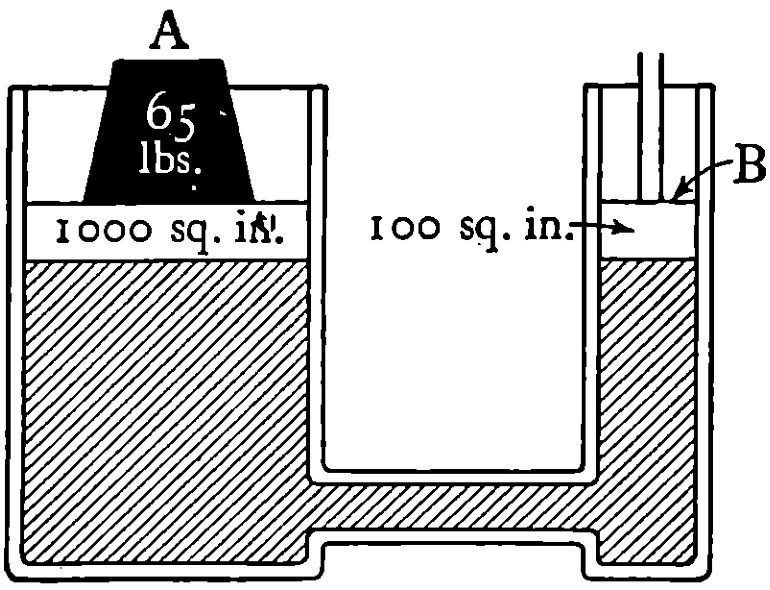
FIGURE 1



a The water does not rise to fill the tumbler completely because of the. . . . . of the air in the tumbler. *a*

b The fact that the water has risen a short distance into the tumbler as shown at C shows that air is slightly. . . . . under pressure. *b*

FIGURE 20



This diagram represents a hydraulic press.

a If the small piston B has an area of 100 square inches, and the large piston A has an area of 1000 square inches, a force of . . . . . . pounds must be applied at B to lift the 65-pound weight. *a*

b Complete this statement: "A force applied at one point in a liquid is transmitted. . . . . . in all " *b*

Number of blanks correctly filled ——— ÷ 2 = *Score in Part II*———

A few exercises from the *Iowa General Information Test in American History*, and a sampling of sixteen of the items tests comprising the *Columbia Research Bureau Physics Test* are reproduced here as illustrations of the more comprehensive types of high school achievement tests.

IOWA GENERAL INFORMATION TEST IN AMERICAN HISTORY, Form B

1. The war which was fought on land, under the water and in the air was the. . . . . . (1) ——————
2. The canal which was opened for world traffic in August, 1914, was the. (2) ——————
3. New Amsterdam was settled in 1623 by the . . . . (3) ——————
4. The great nations which entered into the struggle to possess the new land of America were England, France, Holland and . . . . (4) ——————
5. The first white men to reach America's shores were trying to find a passage to a country called . . . . (5) ——————
6. The principal occupation of the colonists was . . . . (6) ——————
7. The Fourteenth Amendment defines all persons born or naturalized in the U. S. and subject to the jurisdiction thereof, as . . . . (7) ——————
8. The greatest commander of the South in the Civil War was. . . . . (8) ——————

9. For half a century or more a struggle for colonization supremacy in North America was waged by England and....... (9) ____________
10. The U. S. bought Alaska from the government of .... (10) ____________
96. The first important newspaper to continue for a long period in the U. S. was the.. (96) ____________
97. The Webster-Hayne debate took place in the year .... (97) ____________
98. The laws passed by many state legislatures, from 1870 to 1890, fixing the maximum freight and storage rates and compelling equality of treatment of all shippers, were called .... (98) ____________
99. The first conspicuous participation of the U. S. in world politics was the interference in .... (99) ____________
100. Virginia was commonly known as "The land of the (100) ____________

COLUMBIA RESEARCH BUREAU PHYSICS TEST [10]

1. The liter is one tenth of a cubic meter................ ( )
2. The work done by the large piston of a hydraulic press is frequently greater than that done on the small piston... ( )
3. Ten centimeters is approximately 4 inches............ ( )
4. In any enclosed liquid a force applied to any area of the surface is transmitted undiminished to every equal area in the liquid............................ ( )
5. A change of atmospheric pressure on the surface of a lake has an effect upon the pressure at the bottom......... ( )
6. The pressure due to a depth of 1 meter of water is 100 gr. per sq. cm................................ ( )
7. In a vessel filled with water the pressure at any point on the side is the same as at any point on the bottom...... ( )

PROBLEM. (For statements 8 to 11.) *A rectangular box whose base is* 20 *cm. by* 50 *cm. is set horizontally and filled to a depth of* 10 *cm. with water of density* 1 *gr. per cu. cm.*

8. The weight of water in the box is 10 kilograms........ ( )
9. The water pressure on the bottom of the box can be found by dividing the weight of water by the area of the bottom ( )
10. To double the pressure on the bottom due to the water pour into the box enough water to make the depth 20 cm. ( )

[10] Reproduced with the permission of the publishers, World Book Company, Yonkers, N. Y.

11. The water pressure on the bottom of the box is twice that on the side at the same distance below the water surface ( )

---

140. A copper atom differs from a copper ion in the number of its electrons . . . . ( )
141. The period of an oscillating electrical circuit can be decreased by using a smaller capacitance . . . . ( )
142. It is inconsistent with the principle of conservation of energy to say that heat energy may become unavailable . . ( )
143. Electrons flow through a perfect condenser . . . . ( )
144. An electron moving south in an east-to-west magnetic field will be deflected to the east . . . . ( )

Diagnostic and Analytic Tests. Any test which yields separate measures of highly related abilities underlying achievement in a subject may be called a diagnostic test. Thus a test of general mental abilities is a diagnostic instrument to the extent that it points out or diagnoses the specific weakness or strength of individual pupils. Diagnostic tests are designed to discover particular strengths and weaknesses on the part of the individual child, and within reasonable limits to identify the causes lying back of them. Tests possessing such diagnostic power, while unquestionably of the greatest practical value to the teacher, are very few in number. By far the greater number of standard tests that have been developed thus far are designed to take "cross sections" of general abilities or achievement rather than to isolate specific weaknesses and their causes.

A test to be effectively diagnostic must be focused upon a relatively narrow field of achievement, but it must be comprehensive enough to cover all of the separate but related skills upon which accomplishment in the subject depends. Thus, in general, diagnostic tests tend to confine themselves to specific subject fields such as arithmetic, language, and reading. To afford useful diagnostic information a test must sample into the basic skills underlying success in the subject to such an extent that the results furnish an accurate picture of the status of these abilities. Furthermore, educational diagnosis in the strictest sense implies that underlying causes of deficiencies must be revealed. As a matter of fact the impossibility of accomplishing this is undoubtedly one of

the greatest weaknesses of educational testing. In recent years serious efforts have been made to develop educational measuring instruments with real diagnostic power. Obviously, until the individual pupil's weakness and the underlying causes are revealed the preparation and application of corrective devices is futile. Ultimately, real diagnostic instruments and real remedial material will be made available in the major subject fields. Instead of testing two or three

SPECIMEN OF DIAGNOSTIC TEST IN ARITHMETIC[11]

Standard Mathematical Service

COMPASS DIAGNOSTIC TESTS IN ARITHMETIC

RUCH—KNIGHT—GREENE—STUDEBAKER

EDITED BY G. W. MYERS

TEST I: ADDITION OF WHOLE NUMBERS: FORM A

Name.. Grade........................ Boy or girl?..............

Age.............When is your next birthday?........................How old will you be then?..............

School...Date........................
(Name) (City) (State)

SUMMARY OF PUPIL'S SCORE	PART 1	PART 2	PART 3	PART 4	PART 5	TOTAL
Scores on Parts of Test						
Educational Age Equivalent						
Grade Equivalent of Score						

PART 1—BASIC ADDITION FACTS

Add:

0+2=	4+1=
1+7=	5+3=
2+2=	1+0=
7+6=	6+4=
2+5=	3+4=
2+1=	2+7=
0+9=	1+6=
9+2=	3+3=
3+9=	7+4=
5+8=	0+0=
9+1=	9+9=
8+5=	7+5=
2+8=	9+4=

0	9	5	2	3	1	7	9	8	5	7	6	4	2
2	1	0	3	4	1	5	2	4	4	4	2	1	6

9	4	8	0	4	2	6	5	2	4	3	7	8	8
5	0	6	7	2	5	6	3	4	5	7	3	3	8

9	0	1	9	2	7	1	7	1	7	5	8	3	6
6	9	7	7	8	6	8	8	6	7	9	0	9	3

8+3= 7+2=

Score on Part 1 = Number right =..............
[Total possible score = 70 points]

PART 2—HIGHER DECADE ADDITION

Add:

2	1	6	8	18	4	9	7	24	10	6	16	14	0	4	24	21	4
20	21	22	13	6	12	29	25	7	2	12	6	7	10	42	2	4	10

32	1	4	38	4	33	7	1	11	5	15	0	1	33	40	2	1	5
6	18	16	7	20	3	13	24	9	10	9	12	11	5	2	25	14	28

31	6	9	2	24	19	8	9	32	10	3	7	0	2	6	7	19	9
9	11	20	36	9	4	16	27	2	7	18	26	31	26	13	12	2	23

[11] Copyright 1925, Scott, Foresman and Company.

PART 2—Continued

Add:

19+7=	14+2=	10+3=	20+7=	28+7=	0+20=
12+2=	11+7=	18+8=	26+6=	18+7=	22+ 0=

Score on Part 2 = Number right = __________
[Total possible score = 66 points]

PART 3—COLUMN ADDITION

Add:

4	7	6	9	6	9	9	4	7	8	2	6
7	4	6	4	4	7	2	2	6	0	4	3
5	9	9	6	0	8	4	9	7	1	3	5
	1	3	4	7	9	7	7	4	2	8	7
		4	9	5	1	5	8	8	4	1	0
				0	6	9	8	7	1	8	4
							8	6	2	6	7

Score on Part 3 = Number right × 5 = __________
[Total possible score = 60 points]

PART 4—CARRYING IN COLUMN ADDITION

Add:

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
29	98	17	47	117	766	27	9132	7027	162	653		
47	13	13	4	960	982	96	5112	249	4914	5156		
		67	13	517	932	77	2343	7191	9737	7360		
			49		448	84	7417	9235	67	2661		
			29			98		9059	9280	4099		
						17			427	9730		
						26				9222		

12. Copy and add: 62+604+827+797+997 = ? (Put your work under 12 above.)
13. Copy and add: 76, 64, 60, 97, 35, and 20. (Put your work under 13 above.)

Score on Part 4 = Number right × 10 = __________
[Total possible score = 130 points]

PART 5—CHECKING ANSWERS IN ADDITION

Directions: Some of the printed answers below are right, and some are wrong. Check each sum by adding downwards, Write your check answer on the line at the top of the example. The first one is already done correctly.

797							
213	165	6566	9270	7774	887	162	653
128	923	3157	9984	9447	756	4914	5156
456	928	4918	4763	7267	798	9737	7360
787	466	7194	7163	2848	71	67	2661
	2484	77	9650	9192	700	9280	4099
		21792	3083	36528	206	427	9730
					3618	24587	9222
							38861

Score on Part 5 = Number right × 10 = __________
[Total possible score = 70 points]

Copyright, 1925, by Scott, Foresman and Company

times each school year the teacher will test immediately upon the completion of a narrow subject unit. Diagnosis will reveal the specific weaknesses of the individual pupil and the use of properly constructed remedial material will do much to eliminate the defects while there is still time to do something about them. The promptness with which a corrective program is applied has much to do with its effectiveness.

The term "diagnostic" as applied to educational tests has resulted in many misconceptions. Fundamentally, all tests

may be considered diagnostic in the sense that they actually yield specific information about pupil achievement. However, the diagnosis afforded by most present-day high school tests is extremely general. Many so-called diagnostic tests are not diagnostic, but are merely moderately analytical. This is particularly true of most of the tests in such subjects as language, reading, history and science. An illustration may show the distinction which it seems desirable to make between the diagnostic and analytic types. It is unfortunate, however, that the total absence of suitable diagnostic instruments in the high school subjects makes it necessary to choose this illustration from the elementary school field.

Diagnostic power in a test implies its ability to measure one level at a time the related basic skills upon which achievement in the subject depends. For example, in diagnosing a pupil's failure to do acceptable work in arithmetic it may be found that his most immediate difficulty lies in the addition of whole numbers. This is only one of a score of similar possibilities. The diagnosis of his failure in this field is in turn found to involve a rather large number of related abilities. His trouble may be with the basic addition facts, the higher decade facts; it may involve the inability to keep his attention on one thing long enough to add up a long column of figures, or it may be due to his inability to carry from one column to the next. The relation of each of these skills to each other and to the total process in the case of addition of whole numbers is shown clearly in the reproduction of Test 1 of the *Compass Diagnostic Tests in Arithmetic* which appears on pages 33 and 34. The diagnostic procedure here is based on the assumption that mastery of the total process can be no stronger than the weakest link in the chain of related skills. Accordingly, each skill called into play in the total process is so far as possible isolated and measured. For instance, each of the basic facts called for in the examples in *Part* 4, *Carrying in Column Addition* appears as such in *Part* 1, *Basic Addition Facts.* Each higher decade addition fact which appears in the total process as expressed in Part 4 also appears as such in *Part* 2, *Higher Decade Addition.* The column addition difficulties which appear in Part 4 also appear in Part 3 of the diagnostic test.

The net result is that each of the related skills is isolated one at a time until the underlying causes of the pupil's failure to do addition of this type are revealed.

A further illustration of the distinctive features of this more exact type of diagnosis is taken from a typical problem in physics involving long division. The problem itself is as follows:

> How many British Thermal Units (B.T.U.) of heat would be generated in stopping a car whose kinetic energy moving at twenty miles per hour was 37,773 foot-pounds?

In addition to the ability to read and comprehend the problem, and the specific knowledge that one B.T.U. is equivalent to 778 foot-pounds of energy, the student must be able to marshal the following arithmetical skills in solving this problem which involves the division of 37,733 foot-pounds by 778 to secure the required answer of 48.5 B.T.U.:

Addition	Subtraction	Multiplication	Division
28 + 3	3733	4 × 8	
	− 3112		778)3773
		4 × 7	
56 + 6			
	6613	8 × 8	
	− 6224		778)6613
35 + 4		8 × 7	
	3890	5 × 8	778)3890
35 + 3	− 3890		
		5 × 7	

It is thus apparent that instead of this problem being the rather simple one of dividing one number by another, it involves three skills in bringing down numbers in subtraction in long division not shown here, plus the proper functioning of sixteen specific numerical skills, the failure of any one of which may cause the results to be incorrect.

In contrast with the specific diagnosis which appears to be possible in the case of arithmetic is the general type of analysis which seems to climax the best efforts of test makers in the fields of language, reading, and certain other highly complex subjects. Attempts to analyze language and read-

ing with a view to the construction of diagnostic instruments immediately encounter the impossibility of relating to each other in any casual sense the several phases of the subject upon which achievement in it seems to depend. For example, Sangren and Woody in their analysis of the underlying silent reading abilities upon which the *Sangren-Woody Reading Test*[12] is based do not establish any causal relationship between such factors in silent reading ability as word meaning, rate of reading, comprehension of fact material, comprehension of central thought, ability to comprehend and follow directions or ability to comprehend and organize material read. Measures of each of these abilities are necessarily treated as independent and unrelated aspects of total silent reading ability. Low scores on these phases of reading ability do not reveal causes; they merely reveal a condition. This is not a criticism of the *Sangren-Woody Reading Test*; it is merely a statement of the general characteristics of tests of this type.

Possibly a more striking example of analysis without diagnosis is found in the case of a number of recent attempts to measure elementary school language abilities. The *Pressey Diagnostic Tests in English Composition* really measure four aspects of language ability, ability to capitalize, ability to punctuate, ability to identify certain correct and incorrect usages, and ability to recognize sentences. No causal effects of one ability on the other or upon the total ability in language are established. Thus, this test in common with many other so-called diagnostic tests is in reality only a reasonably good survey or analytical test. The same may be said of the much more comprehensive *Iowa Elementary Language Tests*[13] which measure eleven specific aspects of language ability. However, the authors in this case do not claim diagnostic power for these tests, but classify them merely as analytical instruments.

An illustration may be an efficient way of summarizing the essential features of diagnostic and analytical tests. On certain of the points along the rim of the Grand Canyon there are look-out stations equipped with telescopes, each

[12] World Book Company, Copyright, 1927.
[13] Educational Test Bureau. Copyright, 1929.

pointed and focused upon a specific spot of beauty or grandeur. From each of these a separate view of the beauty spots of the canyon is secured. From the composite of all of these views a much more accurate appreciation of the total panorama is obtained, yet each view is quite independent of each other one. This is typical of the way in which tests of the analytical type operate. In distinct contrast with this example the best illustration of the operation of the diagnostic type of test is to liken it to an inverted pyramid made of bricks. The removal or the crumbling of a single brick at any point in the wall will cause it to fall. The accompanying diagram (Figure 1) may be helpful in clarifying the essential differences in these two types of tests.

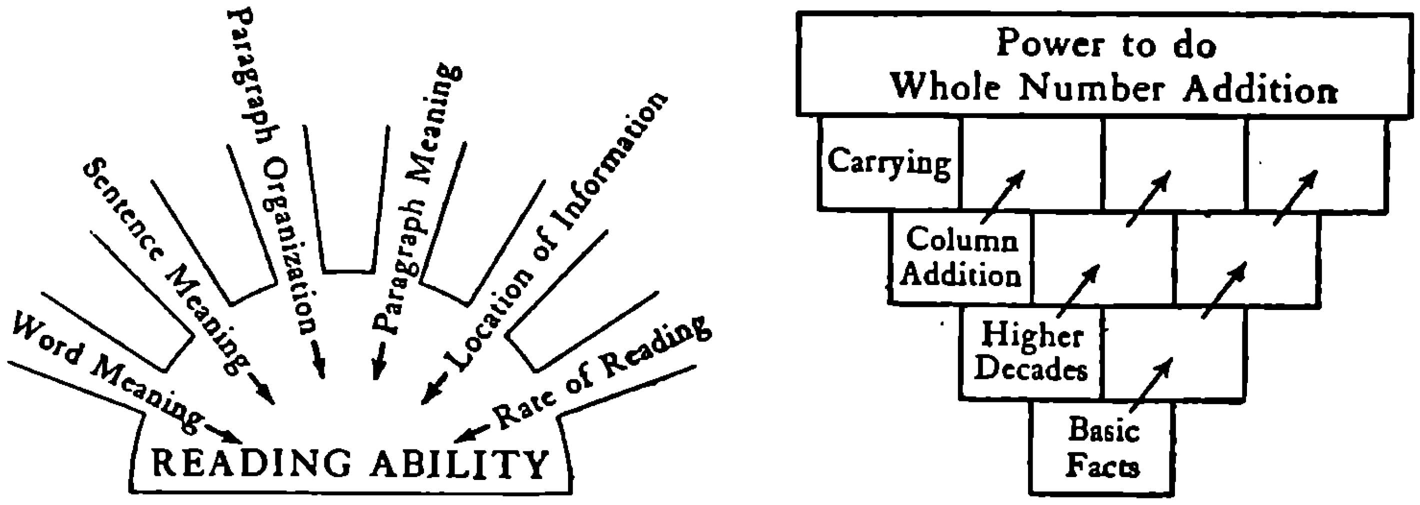


Figure 1. Analysis and Diagnosis

Discussion Exercises

1. Enumerate as many as possible of the qualities in which standard tests are similar to informal types of examinations.
2. Enumerate as many as possible of the qualities that distinguish standard tests from informal examinations and tests.
3. Specify six things which educational tests, when properly used, do for the classroom teacher and for his pupils.
4. What are the features of tests which distinguish them from scales?
5. How can a *test* be made from a *scale?* Illustrate.
6. Distinguish between general ability tests and tests of achievement.
7. What does the process of standardizing a test involve?
8. In what ways are prognostic tests different from general ability tests?
9. In the field of achievement tests what qualities distinguish *rate* tests from *power* tests?

10. In your opinion which type of tests, rate or power, or a combination of the two, represents the truest measure of ability? Why?
11. Explain how Quality Scales actually operate to aid the teacher in the more accurate measurement of the particular achievement, as handwriting, or English composition.
12. Which kinds of achievement test (survey, diagnostic, analytical) are of most value to the classroom teacher? Defend your answer.
13. What is the principle of construction on which the *Compass Diagnostic Tests in Arithmetic* are built?
14. Under what conditions would it be possible to develop such types of diagnostic tests in other fields? Which fields, for example?

Selected References

Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching.* Boston : Houghton Mifflin Company, 1931.

Kelley, T. L., *Interpretation of Educational Measurements.* Yonkers, New York : World Book Company, 1927.

Lang, A. R., *Modern Methods in Written Examinations.* Boston : Houghton Mifflin Company, 1930.

Madsen, I. N., *Educational Measurement in the Elementary Grades.* Yonkers, New York : World Book Company, 1930.

Odell, C. W., *Educational Measurement in High School.* New York : The Century Company, 1930.

Odell, C. W., *Traditional Examinations and New-Type Tests.* New York : The Century Company, 1928.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago : Scott, Foresman and Company, 1929.

Ruch, G. M., and Stoddard, G. D., *Tests and Measurements in High School Instruction.* Yonkers, New York : World Book Company, 1927.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York : Silver, Burdett and Company, 1928.

Webb, L. W., and Shotwell, Anna M., *Standard Tests in the Elementary School.* New York : Ray Long and Richard Smith, Inc., 1932.

Wilson, G. M., and Hoke, K. J., *How To Measure* (Revised). New York : The Macmillan Company, 1928.

Woody, Clifford, and Sangren, P. V., *Administration of the Testing Program.* Yonkers, New York : World Book Company, 1933.

CHAPTER III

THE IMPROVEMENT OF TEACHER-MADE MEASURES OF ACHIEVEMENT

THIS chapter summarizes the following points involved in the general improvement of teacher-made measures of achievement:

a. Need for clarifying the aims and purposes of examinations.
b. Educational functions served by examinations.
c. Need for improving the different types of examinations.
d. Recognized limitations of the informal objective test.
e. Importance of essay-type test for certain types of classroom use.
f. Improvement in the teacher-made essay-type examination.

I. EXAMINATIONS AND THE CLASSROOM TEACHER

19. **Teacher Evaluation of Instruction.** Teachers, practically since the beginning of organized education, have made use of some form of examination as a means of evaluating the end-product of their teaching effort. It was not until the beginning of the present century, however, that measurement of educational achievement began to be taken quite seriously. Today measurement of some kind has come to be considered an indispensable part of effective teaching practice.

Examinations have an ancient origin. Where and when they first appeared is not known. Oral examinations are probably the oldest form, evidences of which are found in ancient literature. In the Old Testament we are told the "Gileadites, at the passage of the Jordan tested the ability of the Ephraimites to pronounce the word 'Shibboleth.' Any Ephraimite who failed to pronounce the aspirate and answered 'Sibboleth' was slain on the spot. On that fatal day forty-two thousand are reported to have failed in their

examinations."[1] Socrates, in a method he has made famous, subjected his pupils to exhaustive and searching questioning. Oral quizzing, Socratic or otherwise, has undoubtedly been a part of classroom procedure from the beginnings of teaching activity—in fact, there have been and still are times when it constitutes practically the whole of the teaching act.

20. Shift from Oral to Written Examinations. Written examinations are probably of more recent origin than oral quizzes, but even written examinations date back many centuries. As early as 2200 B.C., China had an elaborate national system of examinations for the purpose of selecting her public officials and these examinations have been known down through the ages for their unusual severity.

In America, written examinations are as old as formal education itself. According to available records the first examinations of note in this country were those of Boston in 1845. One of the specified duties of the Boston school committee was to make an inventory of the schools each year. This annual inspection included an oral examination of all the pupils. As the number of pupils increased, this practice naturally became impossible. It was then decided to quiz only the highest class in each school. This soon became impossible also because of the rapid growth in enrollment, and as a result the examining became quite perfunctory. Finally a sub-committee was appointed to prepare written examinations and administer them to the grammar departments of the Boston schools. These examinations covered the subjects of astronomy, geography, grammar, history, and natural philosophy. The details of preparation, administration, and interpretations of these examinations were carefully worked out to make the examinations as fair as possible. Questions were carefully graded in terms of difficulty, and rules were prepared to guide the committee in the correction of papers. On the basis of these examinations the schools were ranked in order of merit. The committee, apparently, recognized the problem of "immeasurables" as they emphasized the fact that "the examination measured intellectual activity and acquirements

[1] Lang, A. R., *Modern Methods in Written Examinations*, p. 1, Houghton Mifflin Company, 1930.

only, and that traits of conscience, respect for order, religious sense, and the like might reverse the ranking of the schools." The Report of The Grammar School Committee is reprinted in full by Caldwell and Courtis.[2] A careful reading of this report reveals the completeness with which they carried on their examination program.

21. Influence of Horace Mann. This Boston examination project is truly a high light in the history of education in the United States, and it made a real impression upon Horace Mann, who at that time was Secretary of the Massachusetts Board of Education.[3] As editor of the *Common School Journal*, Horace Mann published extracts from the report and made many noteworthy comments[4] on the subject of examinations. He concluded that the new written examination was so superior to the old oral quiz that no school committee would ever again "venture to relapse into the former inadequate and uncertain practice." The reasons advanced by Horace Mann in defense of the written examination were as follows:

1. It is impartial.
2. It is fairer to the scholars.
3. It is more thorough than older forms of examinations.
4. It prevents the "officious interference" of the teacher.
5. It "determines, beyond appeal or gainsaying, whether the pupils have been faithfully and competently taught."
6. It takes away "all possibility of favoritism."
7. It makes the information obtained available to all.
8. It enables all to appraise the ease or difficulty of the question.

However much we may wish to disagree with the accuracy of this characterization of the written examination, it must be admitted that his ideals were apparently those represented by modern tests, but his instruments were inadequate. It is significant to note that in successive issues of the *Common School Journal* Mann suggested most of the elements in examinations which are found in the modern measurement movement. It remained, however, for such men as Rice,

[2] Caldwell, O. W. and Courtis, S. A., *Then and Now in Education*, 1845-1923. Yonkers, New York: World Book Company, 1923.

[3] Horace Mann, no doubt, exerted quite an influence on this group and much of its point of view on examinations was probably a reflection of his ideas.

[4] Caldwell and Courtis, *op. cit.*

Thorndike, Courtis, Stone, and others to stimulate the development of the objective examination and the standardized test as they are known today.

II. Functions of Classroom Measurement

22. Clarification of the Aims and Purposes of Examinations. Teachers and students of educational administration and supervision are not at all in agreement as to the place of examinations in the educational program. Some of the criticism is levelled at the continued use of examinations as a means of evaluating the results of instruction. More of it apparently is directed at the types of examinations used. There is little question but that on the whole this critical attitude has had a good effect on the attitude of students, parents, and teachers toward examinations. Many significant improvements in examination methods have resulted from this critical attitude. In fact, the struggle of the classroom teacher to secure more accurate methods of evaluating achievement than those available at the moment accounts for much of the improvement in measuring techniques which has taken place in the past twenty years. The effect has been wholesome also in that existing notions of the functions of examinations have been clarified so that the three following claims emerge as basic:

1. They provide an accurate basis for the measurement of educational accomplishment.
2. They serve to motivate pupil learning.
3. They give the pupil training in organized thinking and in the systematic expression of thought.

Examinations as Measures of Accomplishment. The primary purpose of any examination is to determine how far individuals have profited from the instruction they have received. Experience with oral quizzes and written tests of the discussion or essay type has convinced teachers that this is a legitimate function of examinations, but that the instruments they are able to build do not exactly live up to their requirements. The oral quiz is unduly time consuming, and furthermore, there are not many individuals qualified by either knowledge or experience to use it. Critical inspec-

tion of the written type of examination reveals the fact that it too is seriously lacking in fairness to the pupil. It is not wholly impartial. It measures many things other than sheer pupil accomplishment. Only a narrow field of the subject matter is sampled. The personal bias of the teacher unduly affects the pupil's score. The recent realization of these and many other limitations of the traditional written examination has resulted in much improvement in the techniques which the teacher of today may use in the evaluation of pupil achievement.

Examinations as Motivators of Pupil Learning. The importance of the motivating function of examinations is obvious, although very little can be said in defense of this value of examinations as used in the past. The older written examinations were generally given at the close of the school term to determine the pupil's grade in the subject. In the absence of stated goals of instruction and learning, and of faith in the accuracy with which the individual and comparative abilities of pupils might be revealed, there was little of motivating value in them for either teacher or pupil. However, it must be recognized that as instructional outcomes become more definitely identified the quality of examinations based upon them can be improved. Without doubt properly constructed examinations do possess motivating power, and certainly their use for this purpose is legitimate. Pupils and teacher alike work best and accomplish most when they have a clear notion of the goals to be achieved and the progress they are making toward their attainment. In this way examinations serve as an effective means of providing to both pupil and teacher the satisfaction and stimulation which comes from a knowledge of their weaknesses, strengths, and accomplishments. The results from examinations also provide the basis for the use of other useful motivating devices such as charts and graphs showing progress. This type of information is frequently found to be most effective in stimulating the pupil to rival his own past efforts.[5]

[5] In many quarters this notion of a pupil competing with his own past efforts is gaining favor over the former practice of comparing the pupil's accomplishment with the published norms of standardized tests. One reason for the popularity of this

Training in Thinking and in Systematic Expression of Thought. In theory, examinations should contribute much to the educative process because they call upon the student to organize systematically his facts and express his ideas about the subject. No doubt good does come to the student from any situation in which thinking actually takes place. However, extensive evidence based upon the examination and analysis of many examination papers written by students in all levels of ability indicates that the expected benefits arising from the training thus afforded in thinking about the subject matter and in expressing these thoughts in definite form appears with disheartening rarity. Such an important educational outcome can not be left to mere chance. In general, the conditions under which a pupil writes his answers to a discussion-type examination are not particularly conducive to careful thought organization. He must work under a certain amount of time pressure. He is prompted to dash off his ideas about the topic as fast as they come to him without much critical selection or organization. The net result in many cases is that the student, rather than building up habits of critical evaluation, selection, and comparison of the various phases of his knowledge about the subject, is led into the formation of bad habits of work. Thus in a way the more he is forced to respond to such situations the more he is encouraged to build up these undesirable habits.

This emphasis on the importance of providing situations which force the student to think and to express his thoughts is not to be interpreted as a direct indictment of the traditional or discussion-type examination. It is rather a criticism of the way in which it is now used. If it is important that the pupil have training in organized thinking and thought expression in the different fields of subject-matter content to which he is exposed, then the conditions should be set up so that this training can be effectively given. This means not the less frequent use of the discussion-type examination but rather its use under conditions that give the student a

idea is probably found in the fact that poor students are frequently able to surpass their own previous performances, but are usually unable to excel the norms of the tests with which their accomplishment is compared.

real opportunity in which to do his thinking about it. Rather than give him an hour or a two-hour period in which to accomplish this, give him two days or two weeks, and then insist that the final results show evidences of critical evaluation, selection, and organization. Insist that it be written in good form and free from common language errors. Very few prize essays have been found among the examination papers written under the conditions imposed in the typical classroom.

23. **Significance Attached to Examinations.** It requires but a casual inspection of educational practices to discover the significance which is attached to examinations by the school, as well as by the teacher, the pupil, and the parent. Pupils spend a great deal of time in preparing for and writing examinations. The school spends considerable time and money setting up an organization for the preparation and administration of examinations. Teachers devote much time to the preparation, scoring, and grading of examination papers, while parents in general set far too much store by the marks earned by their children on school examinations. Thus, in many respects, examinations have come to assume a place of importance in the educational program which is not entirely in keeping with their true values.

Examinations play an important part in the public relations contacts of the school. To a certain extent they carry to the parents in the community the educational purposes of the school, the aims of specific subjects and courses, and the various emphases held important by the instructional agents of their school. Examinations in part serve as a means of revealing to both parent and pupil the basis for a pupil's scholastic rating, his promotions, failures, conditions, awards, and preparation for further educational work.

For the teacher, examinations make possible the setting up of specific objectives and provide a means of determining his efficiency in achieving them. They aid in revealing overemphasis or wrong emphasis in teaching method and make possible the experimental evaluation of subject-matter organization. Furthermore, they may point the way for specific corrective treatment because of individual and class weaknesses revealed by their use.

III. The Improvement of the Teacher's Classroom Testing Techniques

24. Need for Improvement in Classroom Measurement. Even a superficial observation of typical examination procedures in the classroom makes it apparent that the three basic functions of examinations, measurement of accomplishment, motivation of pupil learning, and provision of adequate training in thought and language expression, are not realized in many instances. On the other hand the more formal objective examinations and standardized tests are inadequate in many respects. An objective examination or a standard test which would measure achievement satisfactorily in one school might prove to be inadequate in another because of differences in the emphasis and point of view which have grown out of the specific needs of the locality. Then too, many teachers feel that tests of the objective type tend to place too much emphasis on the mere acquisition of facts by their pupils. This emphasis, in their judgment, motivates a wrong type of learning. Students too often come to think of their work as mainly that of amassing facts and formulas. Written examinations calling for application of facts in the solution of real problems tend to counteract this attitude and to sensitize students to the notion that the true object of education is only in part the acquisition of bare knowledge much of which is temporary and will vanish. Of much greater importance is the ability to deal with facts. For these reasons written examinations prepared by local teachers or other local instructional agents will undoubtedly always be needed to meet the demands for complete and valid measurement of educational achievement. It is important therefore that steps be taken to recognize the importance of these instruments and to make them as effective as possible in the classroom.

25. Improving the Teacher-Made Essay Examination. For many years the essay-type test has been subjected to intense criticism from a number of angles. In spite of these attacks, however, it is still in use in numerous classrooms and doubtless performs a worthwhile function there. While it is true that when the essay-test is subjected to a critical ap-

praisal under research conditions many of the claims which have been advanced for it do not stand up any too well, it is also true that it performs certain functions in the classroom and for the student which the other more objective forms of tests fail to do. The essay-type test is firmly fixed in educational practice. It is a type of examination with which all teachers are familiar, and with all of its faults it undoubtedly possesses sufficient merit to warrant some attention to its improvement.

Possibly the best way to approach the problem of improving the written examination is to examine some of its more outstanding inadequacies. From the point of view of accuracy in measuring and in revealing differences in pupil achievements the traditional examination is limited by its very nature. The fact that the pupil is required to place on record an elaborate and extensive reaction means at the outset that the limitations of time and sheer physical strength make it impossible for him to give many such reactions. Any single examination obviously can not hold the pupil directly responsible for an understanding of, or for the ability to use, or even for a verbal learning of all of the specific informations, skills, ideas, relationships, and generalizations which constitute a given field of subject matter. Moreover, most schools would find it impractical to devote several hours or perhaps days to the administration of a single test. Accordingly, short tests are the rule. It therefore becomes the responsibility of the maker of the examination to select the essentials for testing from the total field. Thus the items comprising any examination represent only a very restricted sampling selected from all of the subject matter involved. Viewed from the constructive side this implies that since restricted samplings are necessary, it becomes more and more important that each item included in this sampling shall contribute as much as possible to the exactness with which the test measures achievement. Some evidence on this point[6] indicates that when particular care is shown in the selection of subject-matter content and in the formulation of the test questions, examinations of the traditional or discussion-type may give measures which are just as high in validity as

[6] Ruch, G. M., *The Objective or New-Type Examination*, 1929.

would result from objective tests made up of items similarly selected.

Accuracy in the measurement of pupil achievement by means of essay-type examinations is further limited by undue variations in the marks or scores assigned to the test papers by the teacher. Unfortunately from the point of view of improving the accuracy of scoring the teacher-made examination this limitation appears to be largely inherent in the type of examination. In general the variation in assignment of marks to teacher-made examinations of the discussion type may be accounted for by:

1. The lack of uniform standards of expectancy, as between school systems, schools in the same system, or the same grades in the same system, and between teachers of the same subjects or grades. This is clearly illustrated by the data in Table 1. This tabulation shows the shifts in standards which enter into teachers' estimates of school products. Let us assume that a given school product such as a composition has a rating-scale value of 64. To the seventh grade teacher of English such a product would represent distinctly superior work for the grade and a mark of "A" would be assigned to it. To the twelfth grade teacher, however, the specimen would appear to be very inferior as a twelfth grade product and an inferior grade (D) would be assigned.

TABLE 1

SHIFTING STANDARDS OF EXPECTANCY

Quality of Product	Grade 7	8	9	10	11	12
72	A	A	A	B	C	C
68	A	A	B	C	C	D
64	A	B	C	C	D	D
60	B	C	C	D	D	F
55	C	C	D	D	F	F
50	C	D	D	F	F	F

2. The unreliability of measurement due to (a) lack of an objective unit of measurement, and (b) a limited sampling of the subject-matter

field. The effect of the lack of objectivity in the unit of measurement may be demonstrated to anyone who will try to measure the length of a table top by using a rubber band as the unit of measurement. The length of the table in rubber-band units depends upon how much tension is placed on the rubber band. See Figure 2 on page 63, for an example of the effect of sampling on the accuracy of measurement of achievement.

The establishment of uniform standards in the teacher's notion of acceptable achievement is probably a human impossibility. The remedy lies not in the attempt to produce it but in placing in the teacher's hands a tangible unit of measurement. Thus it is apparent that, in general, accuracy of measurement of pupil achievement will be increased by using objective forms of the test question and by greatly extending the sampling of items covered in the examination. In a way the one is dependent upon the other. Part of the solution of the problem lies in recognizing the fact that examination exercises calling for extensive discussion and vague answers are not exactly suited to the accurate measurement of the factual aspects of a course. This does not deny the real utility of the discussion-type of examination for certain other classroom purposes, however. If properly constructed and administered, it still provides a most satisfactory supplement to the more factual measures by giving the pupil a chance to reveal how well he is able to think his way through the problems involved in the organization of these facts in relation to the whole subject. This, in itself, is an accomplishment of unquestioned importance.

Advocates of the older type of examination insist that the discussion-type questions have value not possessed by the informal objective tests in that they call for comparison and interpretation of facts, for criticism and defense of opinion. Essay questions allow for some range of choice which makes possible the meeting of differences in courses and readings pursued. The purpose of the written test is primarily to ascertain, not if the student knows this or that in particular, or has done routine work well, but if he knows accurately a considerable amount and understands a considerable variety of matters in terms of their interacting relationships. What is sought is a measure of accurate knowledge of fact, under-

standing of difficult ideas, and reflection based on the ability to interpret and criticize and decide. In short, the questions are devised to test the pupil's ability to make use of knowledge.

Three specific steps appear to be necessary in bringing about improvements in the teacher-made examination of the discussion type. Each of these has been suggested in the previous discussion in this chapter but has not been actually stated.

1. *The exact purpose of the examination must be understood by the teacher as well as the pupil.* Examinations may emphasize either information, skills, or capacity for thought and expression. The emphasis of the essay-type test is definitely on the latter purpose. The main concern is with topics which involve interest-centers or relationships and problematical issues. The factual aspects of the subject can be measured more effectively through the use of objective testing techniques. Questions involving judgments, synthesis, and generalizations are admittedly harder to evaluate because they are frequently controversial in nature, but it is also true that they provide a valuable training to the student, as well as reveal aspects of the pupil's mastery and mind-quality not indicated otherwise.

   Essay-type questions have been generally open to the criticism that they were hastily and carelessly prepared. The advocates of the improved essay examination are quite positive in their insistence that the preparation and selection of suitable essay-type questions should consume at least as much time as is required to grade the answers. If this is done the value and the accuracy of the measurement obtained is almost certain to be increased and a most important opportunity for training the student in thought organization and expression will have been provided.
2. *The content of the examination should be governed by its purpose.* In general, an examination should parallel the subject-matter objectives of the course. If the examination is designed to check mastery of details and facts, materials of that type should be selected and stated in objective form, since the objective examination accomplishes this type of measurement more effectively. The essay-type examination should confine itself to legitimate purposes as stated above.

   The accompanying tentative score card for rating essay-type questions is suggested as a possible means of improving the teacher-made examination exercise by calling attention to the desirable qualities in such questions. Unless a question rates "Yes" on at least seven of the ten items it is certainly of doubtful value and should probably be rewritten or completely eliminated from the examination.

TENTATIVE SCORE CARD FOR RATING ESSAY-TYPE EXAMINATION QUESTIONS

	Yes	Slightly	No
1. Is the question concerned with important phases of the subject?			
2. If the question emphasizes minor details, are they useful in linking up other facts, ideas, theories, involved in the subject? . .			
3. Does the question give emphasis to evaluation and to relational thinking? . . .			
4. Is the question apparently of a suitable degree of difficulty in relation to the other questions in the test?			
5. Is the question stated in such a way as to stimulate thought, to challenge interest of pupils?			
6. Does the question force the student to integrate his ideas around certain interest-centers?			
7. Is the question stated in such form as to force the pupil to sample widely into his background of fact?			
8. Does the question call for any originality of thought organization and expression?			
9. Does the question call for the student to integrate facts gained from different sources?			
10. Is the question limited sufficiently that the pupil has some chance of writing what he really knows about it in a reasonable time?			

3. *Definite rules should be formulated which will so far as possible control the irrelevant factors in scoring the papers.* In the numerous investigations of the essay-type examination, the evidence indicates that competent judges have been unable to agree on what score should be assigned to the performance of an individual pupil. This is one of the most serious limitations of this type of measuring instrument. In fact, many students of the problem have concluded that since the reliability of scoring such an examination is so low, there is nothing else of value in them. On the other hand, other investigators have taken the attitude that an examination to be useful should not only measure the subject matter or the abilities it is

supposed to measure, but that it must be so constructed that it will call forth a consistent performance on the part of the pupil. In addition to these qualities, it must be possible for qualified judges to agree with some consistency as to the quality of the pupil's performance. There is some evidence that thoughtful care in the selection and statement of questions for essay examinations may result in securing a satisfactorily high validity. Sims[7] presented evidence to the effect that the relation between an essay examination and an objective examination over the same subject matter was almost as high as the consistency of measurement of the two tests would permit. (For those who have some statistical background it may be stated that r in this case was .70. The meaning of this expression of relationship is explained in detail in Chapter VIII of this book.) Sims found also that an essay examination graded by a rating method had an average coefficient of objectivity of .77. From a summary of much statistical evidence on this point Ruch concludes that the reliability of the essay examination is not far from .60 to .65, while Sims found a reliability of .72 for his rating method of scoring such examinations.

Many different plans have been proposed for increasing the objectivity of scoring essay examination papers. Osburn[8] found that by providing the scorer with a list of acceptable answers to each question in addition to the use of marked care in the statement of the questions, and a control in the amount of penalty exacted for formal errors, the subjectivity of scoring the essay examination was significantly decreased. Sims, in an investigation mentioned earlier in this section, proposed a rating method of scoring essay examinations. His readers were instructed to use the textbook upon which the examination was based to work out for themselves acceptable answers to the questions and then to grade the questions according to the following directions presented here because of the suggestions they offer for the more objective grading of essay examination questions:

a. "Quickly read through the papers and on the basis of your opinion of their worth sort them into five groups as follows: (a) very superior papers, (b) superior papers, (c) average papers, (d) inferior papers, (e) very inferior papers. (Remember that in a normal group you would expect to have approximately 10 per cent of very superior and 10 per cent very inferior papers, 20 per cent of superior and 20 per cent of inferior papers, and 40 per cent of average papers. Do not, however, try to conform rigidly to this rule. Your group may not be a normal one.)

[7] Sims, Verner Martin, "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating," *Journal of Educational Research*, Vol. XXIV, No. 3, October, 1931.

[8] Osburn, W. J., "Testing Thinking," *Journal of Educational Research*, Vol. 27, No. 6; February, 1934.

b. "Re-read the papers in each group and shift any that you feel have been misplaced.
c. "Make no attempt to give numerical grades or to evaluate each question. Place each paper on the basis of your general impression of the total.
d. "Assign letter grades to each group ; beginning with A for the very superior group, B for the superior group, etc." ·

TABLE 2

COMPARISON OF GRADES ASSIGNED AN EIGHTH GRADE ARITHMETIC PAPER WITH AND WITHOUT THE USE OF SCORING RULES
(Adapted from Fauber)

Marks Assigned	Mid-Points	Without Rules	With Rules
82.5 – 87.5	85	1	
77.5 – 82.5	80	0	
72.5 – 77.5	75	2	2
67.5 – 72.5	70	6	17
62.5 – 67.5	65	5	13
57.5 – 62.5	60	10	12
52.5 – 57.5	55	5	3
47.5 – 52.5	50	1	
42.5 – 47.5	45	2	1
37.5 – 42.5	40	2	
32.5 – 37.5	35	2	
27.5 – 32.5	30	0	
22.5 – 27.5	25	2	
17.5 – 22.5	20	1	

Number Teachers	39	48
Average	56.8	64.9
Range	20 – 84	40 – 77
Range of the middle half of the grades	15.5	10.3

Approximately twenty years ago Kelly[9] conducted an investigation into the causes of the variation in teachers' marks assigned to examination papers. He found that the effect of the use of a rather definite set of rules greatly reduced the disagreement of the teachers when the same papers were re-scored. In 1926 Fauber[10] repeated Kelly's experiment using examinations in arithmetic and

[9] Kelly, F. J., *Teachers' Marks*, Teachers College Contributions to Education, No. 66. New York : Columbia University, 1914.

[10] Fauber, O. W., "An Attempt to Bring About Greater Uniformity in Grading Arithmetic Test Papers." Unpublished Master's Thesis, University of Iowa, 1926.

in history. The results of this more recent study serve to substantiate the findings of Kelly's early investigation. The data in Table 2 are adapted from Fauber's tabulations and show the grades assigned the eighth grade arithmetic paper with and without the use of rules for guidance. The thirty-nine teachers who graded this paper without the rules assigned values all of the way from 20 to 84. The average grade assigned was 56.8. In contrast with this the 48 teachers who graded the paper with the rules to guide them assigned scores which ranged from 40 to 77. If one of these marks (the mark of 40) is disregarded the real range is from 55 to 77. The average grade assigned was 64.9. A good indication of the actual spread of the grades assigned is shown by the range of the middle half of the values. For the grades assigned without rules the spread of the middle half of the grade was 15.5. For the grades assigned with the guidance of the rules the spread of the middle half of the cases was 10.3. Thus the results of experiments of this type indicate that the subjectivity of scoring essay examinations may be reduced considerably by the use of special techniques or rules for grading such questions. In spite of this, however, it is generally concluded that this type of examination must be still quite highly subjective.

26. Specific Suggestions for Increasing the Objectivity of Scoring Essay-Type Examinations. The more objectively the essay examination questions can be graded, the less the results will be influenced by the personal judgment or bias of the scorer. The following suggestions have been found valuable for use in correcting essay-type exercises:

1. Essay-type examinations should be scored by the one who makes out the questions. He should know exactly what responses are expected and write them down.
2. Each pupil taking the test should write his name on the back of the test paper and the scorer should disregard the name until the test is scored. This eliminates the subjective factor of being influenced or biased in judgment because of former contacts with the pupil.
3. The scorer should not mark off for misspelled words, sentence structure, paragraphing, poor writing, etc. Similarly, he should not increase the score for excellence in these things. However, such factors may be indicated or checked on the examination. The reason for this lies in the fact that the function of the examination is to measure the pupil's information in a course and not his ability to write or to spell. If it is desirable to test his ability to write, spell, or use correct written English, suitable, valid, and reliable tests should be given for this purpose.
4. Essay-test exercises can be scored most accurately if each item is corrected in all the tests rather than to correct each entire test sepa-

rately. This permits the scorer to concentrate on the answer to a single test exercise, which better enables him to judge the merits of the several pupil responses to the same question.

5. Rate each question on a scale of ten or twenty and then add the ratings on all exercises to get the mark for the paper. This method helps to objectify the score. Thus the score is based on a number of careful judgments rather than on one complex judgment of the entire paper.

Whatever rules are followed, they will necessarily be arbitrary ones and not always defensible. The significant point in the use of rules is that they provide for a reasonable uniformity in the handling of the papers of all the pupils and also furnish a guide for the control of the irrelevant factors that may affect the objectivity of the grading.

A discussion of the characteristic advantages and disadvantages of the informal objective test is reserved for the following chapter.

Discussion Exercises

1. What in your opinion were the chief gains in shifting from the oral quiz to the written examination of the discussion type?
2. State in your own words the real importance of the major functions of examinations.
3. To what extent do you find that examinations really accomplish these three functions effectively in use?
4. In your judgment which is more important, (1) to compare a pupil's achievement with his own past efforts, or (2) to compare his achievement with the norms of achievement based on large numbers of average pupils?
5. How important do you consider the "training-in-organized-thinking-and-thought-expression" function of an examination?
6. What factors largely account for the persistence of the wide variation in marks of achievement assigned to products of similar quality?
7. What are the strategic points at which to begin the attempt to improve the discussion-type examination?
8. In the light of your present information do you believe that the discussion-type examination has a real contribution to make to the more effective measurement of achievement? Why?
9. Why is it desirable to have an essay-type examination paper graded without reference to the author of the paper?

10. Suggest at least five specific devices or procedures for increasing the objectivity of scoring essay-type examinations.

Selected References

Banker, H. J., "Significance of Teachers' Marks." *Journal of Educational Research*, Vol. 16 : 159-171 and 271-284, 1927.

Bolton, F. E., "Do Teachers' Marks Vary as Much as Supposed?" *Education*, Vol. 48 : 23-39, 1927.

Caldwell, O. W. and Courtis, S. A., *Then and Now in Education*. Yonkers, New York : World Book Company, 1923.

Hamilton, E. R., *The Art of Interrogation, Studies in the Principles of Mental Tests and Examinations*. Harcourt, Brace and Company, 1929.

Kelley, T. L., *Interpretation of Educational Measurement*, Chapter 1. Yonkers, New York : World Book Company, 1927.

Lang, A. P., *Modern Methods in Written Examinations*. Houghton Mifflin Company, 1930.

Lowell, A. L., "The Art of Examinations." *Atlantic Monthly*, Vol. 137 : 62, January, 1926.

McAndrew, William, "Our Old Friend Examination." *Proceedings*, National Education Association, 1916, pp. 527-533.

Odell, C. W., *Traditional Examinations and New-Type Tests*, Chapters VII-VIII. New York : The Century Co., 1928.

Opdycke, J. B., "Constructive Examination." *Educational Review*, Vol. 72 : 33-43, 1927.

Osburn, W. J., "Testing Thinking." *Journal of Educational Research*, Vol. 27, No. 6; February, 1934.

Ruch, G. M., *The Objective or New-Type Examination*, Chapters II, III. Chicago : Scott, Foresman and Company, 1929.

Thorndike, E. L., *Measurement of Intelligence*. Bureau of Publications, Teachers College, Columbia University, 1929, p. 400.

Tyler, Ralph W., "What High School Pupils Forget." *Educational Research Bulletin*, Vol. 9 : 490-492, 1930.

CHAPTER IV

CONSTRUCTING AND USING INFORMAL OBJECTIVE EXAMINATIONS IN THE CLASSROOM

THIS chapter deals with certain problems related to the construction and use of informal objective tests in the classroom:

a. Similarities in appearance and use of informal objective tests and standard tests.
b. Advantages and disadvantages of informal objective tests.
c. Principles for use in constructing objective tests.
d. Major types of objective tests.
e. Guessing in objective examinations.

I. CHARACTERISTICS OF INFORMAL OBJECTIVE EXAMINATIONS

27. **Introduction.** The history of the development of educational measurements indicates clearly the general dissatisfaction with traditional examination methods. Early studies such as those of Starch and Elliot, Johnson and Kelly, and more recent investigations such as those reported by Ruch[1] show clearly the unreliability of all types of examinations in which the judgment of the examiner is a factor. Developments contributing to the improvement of measurements in education have followed two main lines: (1) the construction and improvement of standardized test material; and (2) the improvement of the informal non-standardized objective test ("new examination") commonly devised by the classroom teacher. In many respects these two types of measurement are not fundamentally different. Both utilize *samplings* of subject-matter material to stimulate pupil-reactions. In both, the performance is expressed in terms of a *score*. Both make use of exercises which are character-

[1] Ruch, G. M., *The Objective or New-Type Examination.* Scott, Foresman and Company, 1929.

ized as being objective.[2] Yet in spite of these similarities the two types of tests do not seriously overlap in function.

28. The New Examination.[3] Even though standardized instruments for measuring achievement of school children have become very widely used, particularly in the elementary school, the examination set by the teacher still remains the most frequently used means of measuring the achievement of pupils. Although properly constructed standardized educational tests are superior in certain respects to teacher-made examinations, they will probably never displace them for measuring the results of teaching as indicated by pupil attainment. The teacher needs a measuring instrument adapted to a particular course of study, or to the emphasis which has been given to the subject in the teaching of a particular class. The informal objective examination, constructed by the teacher to fit the instruction which the class has been receiving, is the obvious answer.

A comparison of the use of the essay-type examination question and the objective (true-false) type of exercise based on the same material is given in the accompanying illustration of one teacher's attempt to make the test exercises fit a particular item of instruction. An essay-type question designed to force the student to organize his thoughts on a unit of Civics material is given below along with a pupil's answer to this question, rated by a number of teachers at about 70% of a perfect answer.

Q. What are the purposes for which a political party is formed?

A. "The purposes for which a political party is formed are: (1) to establish a platform that will express the views of the members on certain vital questions; (2) to nominate a candidate or candidates for the purpose of running for office. They in turn will uphold the party platform; (3)

[2] A test or measuring instrument is said to be *objective* when different persons using it obtain the same score or appraisal of the pupil's work, or the same person assigns the same score at different times. In other words, objective scoring eliminates differences due to the fact that the standards of different teachers and also those of the same teacher vary at different times. Objectivity in an exercise is secured by making the subject-matter statement conform to certain recognized forms which permit only one type of acceptable response. This makes possible the preparation and use of a scoring key which in itself precludes any serious introduction of personal opinions or prejudices into the results.

[3] The expression "new examination" is commonly used as a collective designation for unstandardized tests and exercises which may be objectively scored and which require little writing on the part of those being tested. They represent a definite movement away from the "essay-type" of examination.

to separate themselves from other parties with whose principles they do not agree ; (4) to establish themselves as a separate organ with certain principles."

The following exercises show how a group of true-false statements based upon the same general subject matter as the above question may be used as an objective-type test. It will be noticed that the pupil answered one of these four exercises incorrectly.

Directions to Pupil : Study each statement below. If a statement is True underline True ; if False underline False. DO NOT GUESS.

1. A political party is formed by a group of people with common political views.	True	False
2. The purpose of a political party is to stimulate interest in political appointments.	True	False
3. Political parties usually cater to the opinions of the minority.	True	False
4. The actual accomplishment of political policies is dependent upon the election of those favoring the policies.	True	False

It may be helpful here for the student to turn to pages 80 to 84 where he may examine illustrations of many different types of objective testing techniques used in the construction of informal examinations.

29. Standardized vs. Non-standardized Objective Tests. Standardized educational tests are structurally not fundamentally different from informal objective examinations in their basic elements. In fact, standardized educational tests are essentially little more than improved and refined objective examinations. Hence, the requirements of objective test construction suggest the criteria by which the written examination may be improved.

In contrast with their similarities from a structural point of view the functions of the standardized test and the informal objective examination over the same material are quite distinct. The standardized test is intended for use in many different school systems and in connection with many different types of courses of study. It must therefore be general as to content. The maker of a standardized test cannot be sure that its content will actually parallel the

instructional emphasis given the subject in the course offered by any individual teacher. Accordingly, the standardized test is useful mainly for general comparisons of school with school, class with class, or city with city. It is not suitable for use in evaluating the accomplishment of pupils in a class under a particular instructor with a specialized instructional emphasis. By the same reasoning, *the standardized test should probably not be used as the basis for the assignment of class marks in any subject.* The informal objective examination, if constructed in accordance with well-recognized principles and incorporating extensive samplings of the subject-matter content actually taught by the teacher, is, on the other hand, a perfectly suitable basis for the assignment of such subject grades. It is quite probable that even though two objective tests, one standardized and one informal, could be made equal in objectivity, length (in terms of number of exercises as well as testing time), and in reliability of measurement, their functional values in the classroom would still be quite unlike, due to unavoidable differences in their subject-matter content alone. Thus, in general, standardized and informal objective tests must be considered as having quite distinct and separate functions, and are not to be used interchangeably.

Many writers and speakers have advocated that objective tests be employed either in conjunction with examinations of the traditional[4] type or in place of such examinations. Although many of those who have done so have based their opinions largely upon theoretical arguments, others have presented experimental data to support their judgments. Among the latter are Ruch,[5] Wood,[6] and Russell,[7] each of whom strongly recommends the use of objective exercises. Another experimenter[8] has concluded that under satisfactory

[4] Term applied to examinations of the essay- or discussion-type which have been in common use for many years.

[5] Ruch, G. M., *The Improvement of the Written Examination.* Scott, Foresman and Company, 1924, p. 193.

———, *An Investigation of Objective Examinations in the Social Studies.* Scott, Foresman and Company, 1926.

[6] Wood, Ben D., *Measurement in Higher Education.* World Book Company, 1923, p. 337.

[7] Russell, Charles, *Classroom Tests.* Ginn & Company, 1926.

[8] Brinkley, Sterling G., *Values of New Types of Examinations in the High School.* Teachers College Contributions to Education, No. 161, Columbia University, 1924, p. 121.

The results obtained by Brinkley are somewhat at variance with the investigations

conditions there is little difference in the general merit of the two kinds of measuring instruments. The bulk of the evidence, however, seems to favor the liberal use of the objective type of test in all situations in which the measurement of pupil mastery of facts is a major issue.

II. Advantages and Disadvantages of the Informal Objective Examination

The foregoing discussion has pointed out the fact that the standardized and the informal objective test are not seriously unlike in the form in which the exercises may be stated. In fact, both types of examinations make use of the same general principles in the formulation of their test exercises. Both the standardized and the non-standardized tests may include enough items to afford a sufficiently consistent measurement. On the other hand, there are a few very distinct differences between the essay-type examination and the teacher-made objective examination. Chief among these factors are (1) the form of the test exercise, (2) the form and nature of the pupil's response, (3) the adequacy of the sampling of subject matter, (4) the consistency of the results of measurement, and (5) the economy of time for both the student and the teacher.

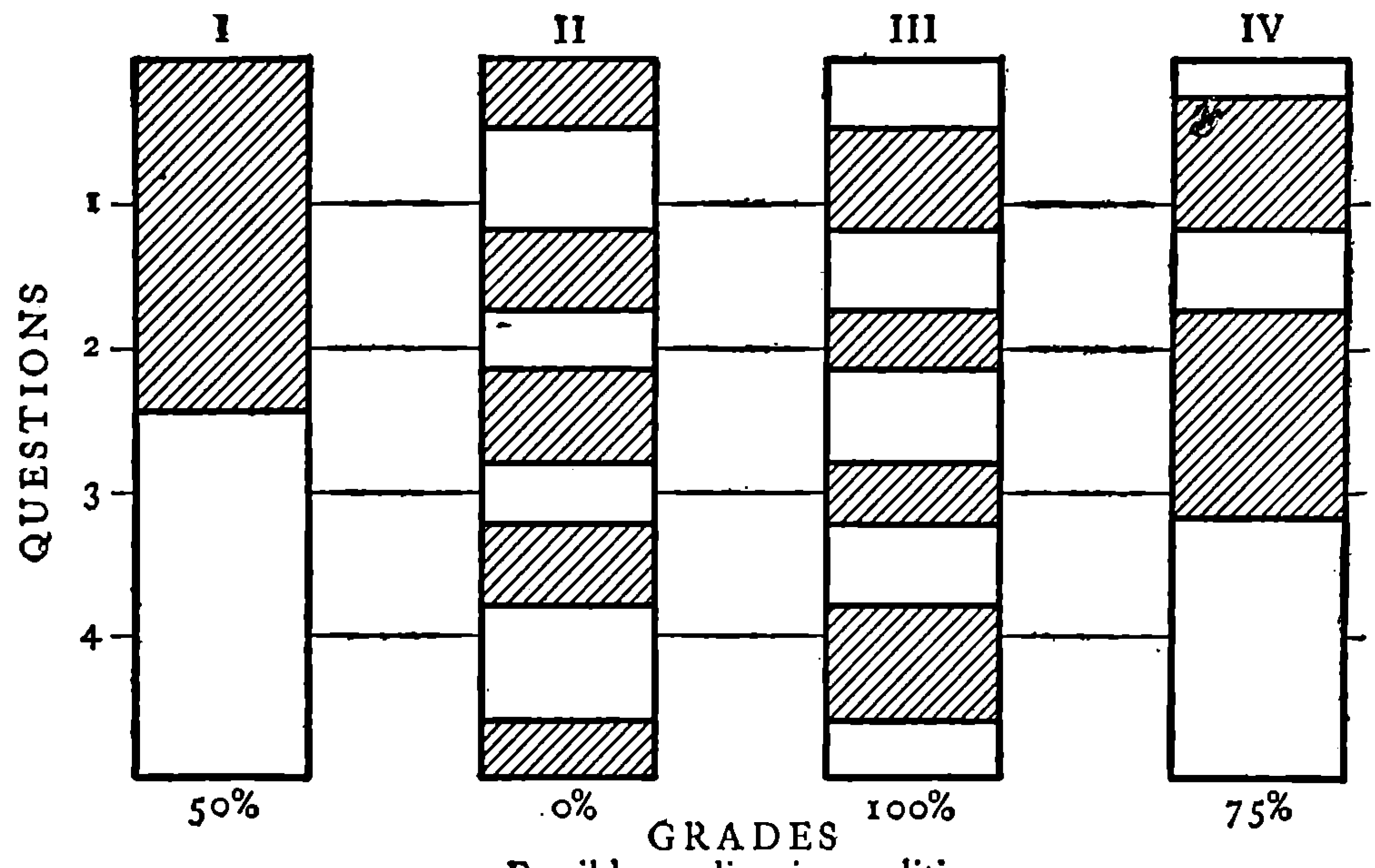
30. Advantages of the Informal Objective Examination. The new type examination has the following significant advantages over the traditional type.

Objective Form of the Stimulus. One of the very significant weaknesses of the essay-type of examination question is the fact that usually it is stated in such a way that it stimulates the student to produce a lengthy discussion which later must be evaluated by the teacher.

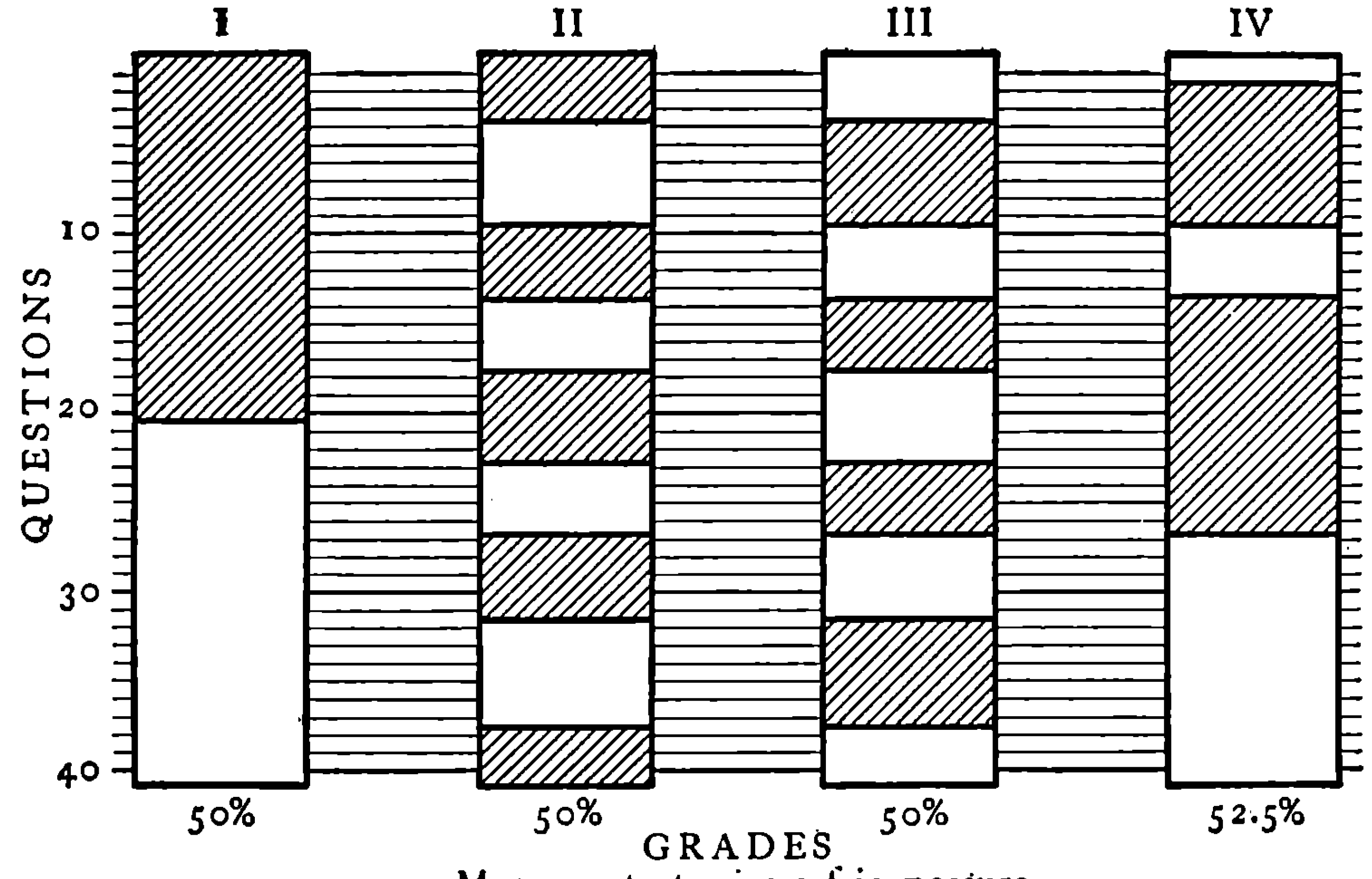
In an objective test the exercises are so stated that the answers are brief, and usually only one correct answer is possible. A perfectly objective test may be scored repeatedly by one person with practically no variation in the scores, or it may be scored by a large number of persons with prac-

of almost all others. This is possibly due to the fact that his teachers were inexperienced in the use of new type examinations and old hands at the use of traditional examinations.

Influence of Sampling on Test Scores



Possible grading inequalities



Many contacts give a fair measure

Figure 2

tically no disagreement in the scores assigned. Thus in the new type of examination the exercises may be evaluated on an impersonal basis, entirely independently of the personal judgment of the examiner. This is true, of course, only when the exercises are constructed in accordance with certain

recognized principles which are discussed in a later section of this chapter.

Brevity of Pupil Response. Fluency of expression and mastery over the language have always been recognized as factors in answering questions in examinations of the discussion type. Because of the nature of the exercises the amount of writing done by pupils in responding to objective tests is reduced to a minimum. This practically eliminates bluffing and the advantage which rapid and fluent writers have over those not so gifted. The fact that one student can write more material than another in the same length of time should not result in his receiving higher marks in his school subjects. Also there is much less opportunity for the teacher to be influenced, either consciously or unconsciously, by good or bad writing, spelling, punctuation, and other similar qualities which have little or no connection with achievement except in the field of English.

Extent of Sampling. The brief response by which a pupil is able to indicate his answer to an objective-type exercise makes it possible for him to respond to a great many more exercises in a given time. Accordingly a much wider range of achievement can be measured or a much larger sampling can be secured than is possible by means of a discussion- or essay-type of examination requiring the same amount of time. This minimizes the chance that pupils may or may not have happened to encounter certain areas of information. The operation of this factor of sampling is shown in Figure 2, on page 63.

This illustration of the unreliability of measurement resulting from improper sampling of subject matter is based on the assumption that four individuals, Pupils I, II, III, and IV have all had the same learning opportunities and have each mastered exactly one-half of the subject-matter of the course. However, in keeping with actual conditions as revealed from classroom data, the particular facts mastered by each individual are not the same throughout. For example, Pupil I, who was perhaps regular in his attendance during the first half of the course, has a mastery of the upper section of the subject matter. This is indicated by the shaded portion of the column. The second individual,

through irregular attendance, spasmodic preparations, or other unknown causes, mastered a few of the facts, missed another section, then learned a few more, etc. Pupil III illustrates the individual who is just as irregular in his mastery as the former pupil but for some reason has learned exactly those items missed by Pupil number II. Pupil IV shows merely another variation of the situation. It might be carried on almost indefinitely but four cases are adequate to illustrate the entire range of variation due to sampling.

Now if a brief essay-type examination consisting of only four questions be sampled from the various areas of the course, it will be noted that distinctly different types of response are secured from these four students, in spite of the fact that each has mastered exactly one-half of the facts in the course. Pupil I knowing the facts in the first part of the work responds to the first two questions with a grade of 50%. The second pupil, by sheer bad luck in the selection of the facts he learned, missed each of the four questions, with a score of zero. Pupil III, through good fortune (or judgment) happens to have mastered the items in the exact areas sampled by the test and thereby makes a perfect score on the test. Pupil IV, as another variation due to chance, scores 75% on the examination. Thus we have a variation of from zero to one hundred per cent on the examination taken by four students who actually have a mastery of exactly fifty per cent of the facts. This type of error in measurement of achievement, which unfortunately is not uncommon, is due to the factor of sampling alone.

The effect of administering an examination made up of many exercises of narrower range is shown in the lower section of the graph. A four question examination permits a range of from zero to one hundred. A forty exercise test permits a range of from 50 per cent to only 52.5%. Thus, it is apparent that the wider sampling gives a distinctly more reliable measure. Clearly the extensive sampling afforded by the use of the objective exercises represents a measure of protection for the student as well as the teacher from unfair and unreliable measurements.

Reliability. The use of objective or near-objective exercises in a test makes it probable that the reliability will be

high. The reliability of a test is the degree to which successive applications of the test yield scores equivalent to those obtained from its first use. In other words, it is a measure of the degree of success in overcoming variable errors in the test scores. Such factors as the disagreement of different persons as to the importance of a certain test exercise, or the correct answer to it obviously affect its reliability. The length of the test in terms of the time required to take it or in terms of the number of items comprising it is one of the significant factors affecting reliability. Reliability is discussed in more detail in Section 63, Chapter VI.

Economy of Time. The form in which the objective exercise is stated makes it possible for the pupil to record his response definitely and briefly. This in turn permits many specific reactions to be called for in a relatively brief period of working time. In this way a much wider area of the subject-matter may be sampled in a given period, resulting in a higher reliability of measurement per unit of working time.

The brevity and conciseness of the pupil's response makes it possible for the scoring of the exercises to be done very accurately and speedily. If the objective examinations are made in accordance with the best practices, the exercises may be scored by simple keys and stencils in the hands of pupils or ordinary clerical help. The fact that it is possible to score and assign grades to a set of 100-item objective-examination papers for a class of thirty-five pupils with a total time expenditure of not over three minutes per pupil is a point which even the most conservative teacher will be glad to consider.

Other minor advantages of the informal objective type of examination might be enumerated. It seems, however, that the five foregoing major items are sufficient to justify the use of objective exercises in preference to those of the traditional type when one desires to measure detailed factual knowledge. There are, however, certain subject matter fields in which the discussion type of examination will necessarily remain the most useful measuring instrument. And even in the fields for which the objective type of test is best suited (informational and content fields) it would be unwise to make it the sole basis of measurement of achievement.

31. Possible Disadvantages of Objective Examinations. A number of rather important criticisms of objective examinations have been brought forward by teachers. The following list, while not complete, probably contains the most significant of these objections.

Neglect of training in thought organization and expression. The fact that the objective examination is so constructed that it makes a minimum demand on the pupil's ability to interrelate the somewhat isolated facts they are taught is in the opinion of many teachers one of the most serious indictments against the objective examination. The nature of the test exercise itself forces this situation. There is no point in attempting to deny the desirability of developing the ability to think through a problem and express a well-formulated judgment. Nor is there any point in insisting that the objective-type test makes any special contribution to pupil growth along these lines. Frankly, it does not, and no such claim is made for it by those who know its qualities best. The best answer to this criticism therefore appears to be to counter-attack by inquiring how much this highly desirable outcome is developed in the essay-type examination. Now, it is embarrassingly true that the amount of such training that is derived from the preparation of an essay-type examination paper is almost exactly zero. Literature has not been greatly enriched by samples from examination papers. Few if any prize essays are produced under the stress of writing feverishly for an hour or more on questions or exercises the limits of which are determined mainly by the writer's imagination. If this training in thought organization is so important, and it is without question a highly desirable product of instruction, then its importance should be recognized and suitable provisions made for its development. The emotional stress under which the student writes his examinations gives him very little chance to think his way carefully through the things he actually knows about the subject. He dashes off his thoughts largely in the order in which they come to him. He has almost no time to consider sentence structure, paragraph organization, or vocabulary choice. The net result is that he forms bad rather than good habits of thought, expression, and work.

It would be much better if a certain proportion of the total time allotted for examination purposes were set aside for the purpose of definitely focusing attention on the development of these skills. The class should be given a list of carefully stated topics related to the subject matter being taught and asked that carefully prepared papers be submitted, after allowing liberal time for preparation. This type of procedure really affords some possibility of giving the student useful training of the type desired, even though it does result in a type of product which can not be evaluated with much objectivity. Perhaps the gain on the one hand is worth the loss on the other, particularly since the use of a comprehensive objective examination will supply the needed measures of achievement so much more economically and reliably.

Objective Examinations Measure Factual Items Only. This objection to the newer type of objective examination naïvely overlooks the fact that almost uniformly essay-type test questions are based on memory and, in the main, test for the factual aspects of the subject. The thought question as a type is not at all inherent in the essay examination. It is unfortunately almost as rare in the essay-type test as it is in the objective form. Furthermore, there is nothing in the objective form which prevents the statement of exercises which stimulate critical and constructive thought. True, most teacher-made tests do not contain this kind of exercise, but that does not mean that they can not be made when teachers become masters of objective techniques and learn to think deeply enough into the subject matter themselves. It may also be worthwhile to point out here that there is a rather definite relationship between a pupil's knowledge of the facts in a situation, and his ability to think clearly in the field. After all, one must have something to manipulate mentally in order to think. A recent study by Kirkpatrick[9] shows that extensive emphasis on objective exercises and testing techniques did not result in sheer memorization of isolated facts to the extent that the pupils were not able to reveal a highly satisfactory mastery and application of general prin-

[9] Kirkpatrick, J. E., *The Motivating Effect of a Specific Type of Testing Program.* University of Iowa Studies in Education, Vol. 9, No. 4, June, 1934, Iowa City.

ciples when tested with a comprehensive examination over the same subject matter. Moreover the high-scoring students on the comprehensive examination designed to measure the ability of the pupils to apply the principles in the course (the mechanics section of high school physics) were also the high-scoring students on the factual aspects of the course when measured objectively. This is believed to be an effective answer to the criticism that the objective examination is innately limited to the measurement of isolated facts and that its use motivates that type of learning exclusively. Without doubt the difficulty lies not so much in objective methods of measurement in general as it does in the kinds of objective material prepared by the individual teacher.

Objective Examinations Are Expensive. Experience in the use of objective examinations indicates that they are most valuable when available for classroom use in printed or mimeographed form. Unquestionably the paper cost is an item of expense which in some school systems may be serious. However, some kind of paper must be used for the examination. Mimeograph paper is approximately as cheap as any. If the teacher is willing to do his own mimeographing or hectographing the extra expense can not be very great, particularly when spread over several classes or terms. As a matter of actual fact, the cost of preparing objective examinations probably represents one of the very minor items of expense in the average school system, when it is considered in terms of the real educational importance of such equipment. Certain types of examinations appear to be affected very slightly by the practice of dictating the exercises and calling for the objective response to be recorded on numbered lines on scratch paper. If care is taken in the administration of an objective test in this way very satisfactory results may be obtained.

It is recognized that these criticisms of the objective examination represent but a partial list. For the teacher who wishes to build up stronger and more differentiated arguments for and against the objective examination, much more elaborate discussions of the problems may be found in a number of the references listed at the conclusion of this chapter.

III. Principles of Constructing and Using Objective Exercises

Before attempting to construct informal objective test exercises the student and classroom teacher who is inexperienced in this type of work should consider the following general principles underlying their construction and use.

32. **Need for Specific Directions.** *Objective test exercises should always be preceded by explicit directions as to just what is to be done.* If the pupils are not familiar with the kind of exercises used in the examination, the direction should also include samples showing just how to do the work. The amount of time allowed should also be stated in the directions if time limits are to be used. Observe this sample set of very satisfactory directions for a true-false examination.

True-False Examination

Directions to the Student: Examine each statement below and decide whether it is true or false. If the statement is true, underline True; if the statement is false, underline False. DO NOT GUESS. This examination will be corrected for guessing. You will be given —— minutes for this examination. Study the sample below. It has been answered correctly.

Sample: Reliability means the extent to which a test measures whatever it does measure. <u>True</u> False

33. **Number of Exercises to Use.** *It is important that a fairly large number of items be included in an examination.* The number of objective exercises to use in a test depends upon two chief factors, (1) the length of the testing period, and (2) the working time required for the pupil to make his response. The length of the testing period is of course determined somewhat by the frequency with which tests are given. If objective tests are given frequently and as a more or less regular part of the daily procedure during the class period as few as twenty-five items may be adequate. If the examinations are given less frequently as, for example, every two weeks or once a month, many more items should be included.[10] Under these conditions one hundred items

[10] The term *item* is used to refer to the smallest unit or part of a test which calls for a distinct answer or pupil response. Thus a single direct recall or yes-no ques-

is probably a minimum. It is important that a fairly large number of items be used in all objective tests. The importance of sampling extensively into the subject matter to be tested is demonstrated clearly by the diagram on page 109. A true and accurate measure of a pupil's information cannot be obtained by brief tests sampling the field by means of a limited number of exercises.

The number of items to be used is also determined to a certain extent by the working time required for the pupil's response. This, in turn, is conditioned by the difficulty of the subject matter and the items covering it. Because of the impossibility of controlling the matter of the innate difficulty of the items it is impractical to attempt to set up any very definite guides as to the number of such items which pupils may answer in a given period of time. In general, recall items are more difficult and require more time for pupil response than do true-false or other recognition forms of objective exercises. An adaptation of data from many sources indicates roughly that for each response to a recall exercise 1.3 responses to multiple response forms may be expected. On the other hand, 1.8 true-false exercises are answered in the time required for one recall item. Ruch[11] reports evidence showing that average pupils answered recall items at the rate of about four per minute, and that upper grade pupils were not markedly slower than high school pupils in answering this form of exercise. This is probably due to the fact that the lower grade pupils omitted a great many more exercises. Experience with the use of true-false and other recognition forms of objective test exercises indicates that where the items are carefully constructed to call for thought from three to four items per minute represent average work for college students. If this may be taken as any indication of the rate of work of such students it is probable that two recall items per minute of a difficulty suitable for use would be adequate. Since the teacher soon learns to adjust the factor of difficulty to keep pace with the increases in

tion, a single pair of expressions in a matching test, a single word to be defined or a single term to be connected with the proper part of a figure is an item. The sample above is one item in a true-false test.

[11] Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929.

ability of the students it is likely that no great increase in the number of responses per minute should be expected from students at the higher grade levels. From the available evidence it is impossible to determine in advance the exact working time required for a given form of objective examination. However, a reasonable estimate may be reached by allowing one minute of working time for each two recall items, each three multiple-response items, and each four true-false items. This means that a one hundred item objective examination will require approximately thirty to fifty minutes of working time depending upon the type and difficulty of item used. For obvious reasons it is undesirable to allow too liberal working periods. It has been found to be quite a satisfactory practice to allow the class to work until by a show of hands it is apparent that approximately ninety per cent of the pupils have had an opportunity to work through the examination at least once.

34. **Selection of the Subject-Matter Items to be Tested.** *Care should be exercised in the selection of the items to compose an objective test.* In the case of a limited field of subject matter in which the aims are quite clear and in which the items of information are specific, practically all of the more important items can be included. In this case, *importance* means that a complete understanding of the subject matter cannot be had without a knowledge of the items referred to. Where the field is broad, the process of sampling, which has been previously discussed, must be resorted to in order to make sure that enough items representing all essential aspects of the subject are included in the examination.

35. **Clear Statement of Exercises.** *Statements should be carefully worded to avoid ambiguity.* The complete elimination of all ambiguity from objective exercises is extremely difficult to achieve. Frequently items require two and even three or more revisions before they can be used. The following exercises illustrate this point:

True-False Exercises

A. Norms and standards are often used synonymously. True False

B. The terms norms and standards are often used synonymously. True False

Exercise *A* creates some doubt in the mind of the reader as to whether it is the terms which are used synonymously or interchangeably, or whether it is the typical scores of pupils (norms) or the desirable goals of achievement (standards) which are used interchangeably. In the second exercise (Exercise *B*) this uncertainty has been removed by the addition of two words. The latter exercise is obviously quite superior to the former.

Objective exercises of the multiple-response type in which a number of possible answers are suggested for choice should be so formulated that the correct answer is not too evident. In other words, the incorrect answers should not be so farfetched or absurd that they can be recognized as such even without any special knowledge of the subject matter being tested.

36. **Range of Difficulty of the Exercises.** *The examination, in each item and also as a whole, should be of such a degree of difficulty that a minimum of either perfect or zero scores will be made.* Test exercises so difficult that no one is able to answer them correctly do not reveal anything, since the examiner can never be certain how much information of difficulty just less than that of the exercises given, the subject does know. The same general condition exists when exercises are so easy that no one fails to answer them correctly. In such a situation there is no way to determine how far above the level of the difficulty of the examination the ability of the class or individual may be. Thus it is a generally accepted principle in educational measurement that test exercises which yield zero and perfect scores are useless and are to be avoided.

37. **Arrangement of the Exercises.** *The items included in an objective examination should be arranged in such a way that there is no regular sequence of answers.* For example, the true and false, or positive and negative, statements in an alternate response type of test should be arranged in random order. The arrangement of exercises in any specific system might become apparent to the student and thus the entire examination would be invalidated. In the following comprehension exercises the answers are arranged in no special system. The grouping and sequence of Yes and No an-

swers is purely a matter of chance which may be likened to the results of tossing a perfectly balanced disc or coin.

SENTENCE COMPREHENSION EXERCISES

May external appearances be deceiving?	Yes	No
Should a person weed out bad habits before they are too firmly established?	Yes	No
May careful observation be of considerable aid in decreasing mistakes in one's conclusions?	Yes	No
Do controversies always result in the satisfactory solution of a problem?	Yes	No
Is it always impossible to misapply talent?	Yes	No
Are all laws enforced with facility?	Yes	No
May there be contention among the members of a parliament?	Yes	No
Are agility and endurance considered good qualifications for an athlete?	Yes	No
Are all festivities characterized by formality and ceremony?	Yes	No
Must exercise be violent to be considered adequate?	Yes	No
Are authorities sometimes quoted in editorials?	Yes	No
Are future events definitely predictable?	Yes	No

38. Arrangement of Answers for Convenience in Scoring. *The directions should require the pupils' responses to be recorded so that scoring is simplified.* For example, the practice of underlining the proper answer, as illustrated, makes it very easy to score the examinations rapidly and without error. The practice of having the pupils record the number of the statement or answer in the case of multiple-choice exercises also simplifies the scoring process.

A satisfactory arrangement of objective items for scoring purposes is shown in the accompanying exercises designed to measure the ability of pupils to use capital letters correctly.

Directions to Pupil: The sentences below are correct, except that certain capital letters are missing from some of the words. *On the line at the right of each exercise, write the word that should begin with a capital letter.* The sample is answered correctly.

Sample: We moved into our new house in march .... March

1. the teacher asked me to help her
2. We saw john when we were in the city
3. We celebrated thanksgiving at home
4. Do you think i should go to school now?
5. Last saturday was a very pleasant day

39. **Arrangement of Items According to Difficulty.** *The items within an examination should be arranged in ascending order of difficulty as soon as possible.* In effecting revisions of objective examinations which have been previously used, the easier items should be placed at or near the beginning of the test with the more difficult items arranged in order of difficulty as revealed by the number of pupil-failures on the items. This information is easily found by making an item count from the previously used examination papers. Of course, in the case of newly made tests which have thus far had no use in the classroom the only basis for such an arrangement of items is the judgment of the maker of the examination.

40. **Elimination of Faulty Exercises.** *Ambiguous and other faulty items which do not discriminate properly between the achievements of pupils with different levels of mastery should be eliminated or revised as soon as possible.* Experience in the formulation of objective exercises indicates that it is very difficult to keep all ambiguous and faulty exercises out of an examination. Unfortunately, it is not always possible to identify these worthless items until after the examination has been used. However, the continued use of examinations containing such items is inexcusable, since it is a relatively simple matter to discover the defective and the useless items in an examination by making an item count of the pupils' responses. Discriminative power[12] on the part of an item is easily determined by dividing the class into two groups on the basis of total accomplishment on the examination and then calculating the per cent of cases in the superior and in the inferior groups who succeed on each test exercise. Items which are missed by the superior group as frequently as they are by the inferior group reveal a lack of power to distinguish between good and poor accomplishment. Examinations which contain many such weak items will fail to discriminate adequately within the class. The student will find a more complete discussion of this problem in Chapter VII.

41. **Use of Various Types of Exercises.** *The principle of variety should be observed.* That is, different types of exer-

[12] See *Glossary* for new terms, *Appendix*, pages 596 to 606.

cises should be combined into one objective examination, especially if the examination is long, in order to avoid the boredom caused by monotony. There are other reasons for using various types of exercises. Teachers often find that some types of important subject matter are adaptable to one form of test exercise and not to another. By using different types of exercises, the teacher can test different kinds of subject matter more adequately. The most important value in the use of different exercises probably lies in the fact that each different form demands a unique type of mental reaction. The results from such testing should be more reliable because of the wider scope of mental abilities which are brought into activity. In an objective test of 100 items in a field such as history it is probably unwise to have all the items of the recall, true-false, or any other one type. It would be much better, for example, to give the students opportunity to respond to 20 recall exercises, 30 true-false exercises, 30 multiple-choice exercises, and 20 matching exercises. Each group of exercises should, of course, be placed together in the test to simplify the pupil's response to the test and to facilitate scoring the paper.

42. **Sources of Objective Test Material.** In making an informal objective test the teacher should make a critical survey of the subject-matter to be tested. He should then make a list of the separate items of information or skills which he desires to test. The number of items will depend upon the purpose and the scope of subject matter to be covered by the test. Only those test items dealing with points of real significance in the course should be included in the test. One of the common errors in the construction of such tests is the inclusion of many unimportant facts found in the reference source which were never intended to be remembered as basic but were merely used for explanatory or illustrative purposes.

By a careful examination of textbooks, teachers' handbooks, course of study outlines, syllabi, and semi-standard objective teaching and testing materials, the teacher should be able to prepare a list of items which would satisfactorily represent the field. To assist the teacher in the selection of the basic concepts to be used in testing in a given field, hand-

books of objective exercises based upon the overlapping content of most frequently used textbooks have been proposed. One such source of objective testing material contains recall, true-false, and multiple-response exercises based on 1500 concepts which are common to the most frequently used textbooks in the field.[13] Teachers should find such material extremely valuable and many more developments along these lines should be encouraged. Such handbooks afford a definite means of determining the real core of the curriculum content for the special subject. They would generally improve the quality of objective testing which goes on in the classroom since each of the desirable types of exercises would be written by an experienced teacher and would also be tried out in advance in the classroom.

IV. Practical Uses of Objective Test Results

43. **Instructional Uses of Objective Tests.** The evaluation of pupil and class achievement is most effectively accomplished through the use of the objective examination. Even if there were standardized tests for the measurement of most of the outcomes of class instruction they would still be unsuited for this type of use. Properly constructed objective examinations within certain limits aid the teacher in determining points at which instructional adjustments must be made. Pupils, likewise, may be led to discover their specific weakness in achievement. One of the most important of all of the instructional values of the objective examination, however, is the use of the resulting test scores as a basis for class marks and other evidences of pupil accomplishment.

44. **Use of Objective Test Scores as Basis for Class Grades.** Pupils' scores from valid and reliable objective examinations afford the teacher's best single basis for measuring and grading pupil achievement within a given subject. The results of objective examinations take out the weakest features of the teacher's marking system. The validity of

[13] Kirkpatrick, J. E., Greene, H. A., *Handbook of Objective Exercises in High School Physics.* Bloomington, Illinois: Public School Publishing Company, 1931. (Thus far only the unit on *Mechanics* has been made available.)

the objective examination depends upon the success of the teacher in formulating objective exercises which really represent the fundamental facts which the pupils have had an opportunity to learn. The reliability of the test is mainly a function of its length. This can easily be controlled. Teachers with a little practice can formulate exercises which are satisfactorily objective. Thus it is apparent that practically the only point where the teacher's judgment needs to enter is in determining the approximate or relative difficulty of the examination as a whole so that the proportionate number of superior and inferior grades may be determined.

Objective test scores may be turned into the particular type of teacher's mark in current use by assigning high grades to a certain proportion of the higher scores, and low or failing grades to a certain proportion of the lower scores. The exact proportion of the class assigned certain grades is not always the same but depends upon the difficulty of the examination and the ability of the class. Sometimes a very easy or very difficult examination or an exceptionally poor or able class may demand a shift in the teacher's marking standard.

The details of the problem of assigning grades on the basis of objective examination scores are so practical and important that they are discussed extensively in Chapters VIII and XXIII.

V. Classification and Illustrations of Types of Objective Exercises

The more common types of objective test exercises may be classified into two main groups: (1) the recall type and (2) the recognition type. In the following pages a classification of the special types of test exercises falling into these two groups is given. Each of the more commonly used types of exercises is later illustrated by one or more exercises complete with directions for use. The varieties of test exercises named and illustrated do not constitute an absolutely complete list, but do include practically all of the types suited for general use. Many slightly different forms can be made by combining or mixing the features of two or more

of the kinds given. Several of the types given illustrate such combinations or mixtures, but many more can be devised by teachers who will study these types of objective exercises.

45. **Classification of Objective Tests.** The main types of objective examinations which have been formulated and widely used to date fall into the following classifications:

I. Recall types
  1. Recall exercises or questions with single correct answers
  2. Completion exercises
    (*a*) simple completion exercises
    (*b*) completion exercises with suggested or controlled answers

II. Recognition types
  1. Multiple response exercises
    a. Multiple-answer exercises
    b. Compound multiple-answer exercises
    c. Multiple-description exercises
  2. Multiple-reason exercises
    a. Multiple-reason exercises with one or more correct answers
  3. True-false exercises
    a. Single and double Yes-No questions
    b. Single and double True-False questions
  4. Incorrect statements
    a. Alternative statements to be corrected
    b. Incorrect statements
  5. Matching exercises
  6. Definitions
  7. Enumeration
  8. Association exercises
  9. Exercises for arrangement in chronological order
  10. Identification exercises
  11. Analogies
  12. Abbreviations or formulae
  13. Classification

46. **Samples of Types of Objective Test Exercises.** The examples given in the following pages have been chosen to demonstrate the manner in which the various types of exercises may be used in different school subjects. In fact, almost any of the kinds illustrated may be used in any one of the commonly taught school subjects. An analysis, however, makes it apparent that some are much better suited for use in certain subjects than in others. It is advisable

in the use of all types of exercises described in the following pages and necessary in the use of most of them that mimeographed or otherwise printed copies of the tests be placed in the hands of the pupils. Even with mature students it has been found that exercises, such as the true-false type, can not be dictated to the pupils without reducing the reliability of the test.[14]

In selecting the samples no effort was made to choose the best possible ones in the various subjects. They are not based on any study of the subject matter commonly taught or the questions commonly asked. In other words, the examples in any one subject, if gathered together, do not constitute a test on the minimum essentials of that subject. So many textbooks and standard tests were consulted in an effort to secure suitable material for these illustrations that specific references to each source are not given.

A. Recall Types

(1) Simple Recall Exercises

This type, which is one of the easiest to prepare and which may be employed in practically all subjects, consists of questions having a single word or expression as the only correct answer. It may, therefore, be made perfectly objective if carefully prepared. The score is the number of correct responses.

Directions: Answer each of the following questions with a single word or expression. Write the word on the line immediately after the last word of the question.

1. Who was the President of the United States during the World War? __________
2. What is the product of x^2, x^3, and x^5? __________
3. What is the dative plural of homo? __________

(2) Completion Exercises

The completion type of exercise is one of the most commonly used of the newer objective forms. From the stand-

[14] Paterson, D. G., *Constructing New-Type Examinations*, World Book Company.

point of pupil response it is probably one of the most valuable tests of this kind. It is very difficult, however, to construct a simple completion test so that the scoring is highly objective. Care should be taken to construct statements which require answers that may be definitely evaluated as either right or wrong. Completion tests are scored by counting the number of statements correctly completed or the number of blanks correctly filled in.

a. Simple Completion Exercises

Directions: The following statements are to be completed by writing one and only one word in each blank.

1. The ———— is a mammal.
2. Water boils at ———— degrees Centigrade.
3. Boyle's Law states the relation between the ———— of a gas and its ————.

b. Completion Exercises with Answers Suggested

Directions: Complete the following statements by inserting one of the words found in the list written on the right of the page. None of the words is to be written more than once.

1. ————	"Snowbound" was written by	1. Bunyan
2. ————	"Pilgrim's Progress" was written by	2. Coleridge
3. ————	"The House of Seven Gables" was written by	3. Whittier
4. ————	"The Ancient Mariner" was written by	4. Hawthorne

B. Recognition Types

(1) Multiple Response Exercises

Exercises of the multiple choice type are among the most satisfactory for use in the construction of objective examinations. With care in constructing and organizing the exercises the scoring can be made very objective. A very satisfactory practice is to number the proposed choices submitted for each exercise. The number of the proper choice can then be written on a line at the right margin of the test

page which very decidedly simplifies the problem of scoring such tests. This is believed to be a distinct improvement over the practice of underlining the correct statement. This type is illustrated in the first of the sample exercises given below.

a. Exercises with only one Correct Answer

Directions: Each of the following statements can be correctly completed by one and only one of the numbered expressions. You are to write the number which stands for the correct expression on the line at the right of the exercise.

1. New York was settled by the (1) Spanish, (2) English, (3) Dutch, (4) French, (5) Portuguese. ________
2. The term of office of a United States Senator is (1) 1, (2) 3, (3) 5, (4) 6, (5) 7 or more years. ________
3. The best temperature for living rooms is about (1) 60°, (2) 68°, (3) 75°, (4) 88°, (5) 78°. ________

b. Compound Multiple-Answer Exercises

Directions: Underline the one expression in each line which is closely related to the expression given in the first line.

1. Woodrow Wilson
   a. Massachusetts, New Jersey, New York, Pennsylvania.
   b. Vice-President, Senator, Ambassador, President.
   c. Republican, Democrat, Prohibitionist, Socialist.
2. Des Moines
   a. Illinois, Indiana, Iowa, Ohio.
   b. (Population) 100,000 ; 75,000 ; 15,000 ; 25,000.

c. Multiple Description Exercises

Directions: Place a check mark (√) before the paragraph which best describes the city of Chicago.

1. Chicago, the largest city of Illinois, is one of the leading livestock markets of this country. It is also noted as a great railroad center. It is situated on Lake Michigan. It is often called the Windy City.
2. Chicago is a city of 3,000,000 population, located in the north central part of Illinois on the Chicago River. Situated in the center of a prosperous agricultural region, it is a leading center in the manufacture of flour and corn products.

3. The city of Chicago, with a population of more than a million, is located on Lake Michigan. Because of its location on the lake and because of its proximity to the mines of the northern states and its transportation facilities, it has become one of the centers of the corn and steel industry.

(2) Multiple Reason Exercises

The characteristics of the multiple reason exercise are practically the same as those of the multiple response type, and on this account samples of these slightly varying techniques are not presented.

(3) True-False Exercises

True-false exercises, because of the ease with which they may be administered and scored, are very popular types of objective test exercises. The two types, those answered by either checking, underlining, or writing the words *true* or *false* after the exercise and those answered by using Yes or No are identical in function although the exercises may be worded somewhat differently. Two main types of true-false or Yes-No exercises are recognized clearly. First, single statements, each covering a distinct item represent the most commonly used form. Second, double true-false statements, one true and one false covering each distinct test item. The first is the more commonly used form. The second is a newer type which has appeared as a result of an effort to devise a test which would automatically correct for the factor of guessing. The evidence shows that when this type of exercise is scored in terms of the number of pairs right the factor of chance is quite largely eliminated. Naturally in making an examination composed of these double true-false exercises the two exercises covering the same item are not allowed to appear close together. Experience shows that it is best to place the paired exercises in opposite sections of the test.

Since true-false exercises are fundamentally the same whether they be of the single or double type, the sample given as an illustration shows a series of paired exercises of the double Yes-No variety.

Directions to the Pupil: Study the samples below. Read each question and answer it by drawing a line under the right answer. DO NOT GUESS.

Samples: Are all people dishonest? YES NO

Are authors often quoted? YES NO

PART A

1. Should all students have the determination for achievement?	YES	NO	1
2. Are all persons true to their best ideals?	YES	NO	2
3. Do people sometimes co-operate for self preservation?	YES	NO	3
4. May disputes ever arise over an expedient alliance?	YES	NO	4
5. Does admiration always involve flattery?	YES	NO	5

PART B

1. Do all students have the same determination to achieve?	YES	NO	1
2. Do many people fail to live up to their possibilities?	YES	NO	2
3. Does co-operation always result in self preservation?	YES	NO	3
4. Do expedient alliances always result in disputes?	YES	NO	4
5. May flattery be an expression of admiration?	YES	NO	5

The remaining types of objective exercises listed on pages 82 and 83 are so infrequently used in the construction of informal examinations of the objective type that they are not illustrated here in detail. Samples of practically all of the varieties of exercises not illustrated may be found in some form by examining any good collection of standardized test material, or by checking through one of the many excellent textbooks on objective testing.

VI. Chance Factors in Objective Tests

47. **Correcting for Guessing.** By the very nature of their construction a great many objective test exercises involve the possibility of a pupil's giving the correct response a certain part of the time by pure chance. Teachers have objected to the use of the new type examinations on this account. The operation of chance is largely excluded in recall type exercises, but is present in practically all of the recognition or

alternate response types. The extent to which it is present depends upon the kind of exercise used. The chance factor can be more definitely controlled in the multiple-response type of informal tests than in the true-false or Yes-No types. In the case of the latter there is, in theory at least, a chance that the pupil may answer at least one-half of the exercises correctly by guessing.

The important question on this point is how much "guessing" actually enters into a pupil's reaction when faced with making a decision on each successive exercise in the test. There are some exercises on which he is absolutely certain of the right response and responds correctly. About other exercises he may be equally certain but makes the wrong response because of misinformation. On still other exercises he may be absolutely without a basis for a response and on such exercises he may hesitate between "guessing" or "omitting" the exercise. In other cases he may not be certain of the correct response, but is urged to respond because of a strong hunch as to the correct answer. In other words, incorrect responses may be due either to answers guessed at, but incorrect, or answers not guessed at but incorrect because of sheer misinformation.

Pupils undoubtedly are often influenced in their responses by some residual knowledge of the answer, and thus do not make purely random guesses. This possibility has brought about the use of two conflicting types of instructions in the administration and the grading of objective tests. One group instructs the pupil to guess when uncertain, the other definitely instructs the pupil not to guess but to omit the answer to the exercise when uncertain as to the correct one. It appears, however, from most recent researches[15] on this point that the scores on objective tests are slightly more valid and reliable when pupils are instructed to omit the answers to exercises when the answering is nothing more than a pure guess. This does not mean that they should not attempt to answer those exercises about which they possess fragmentary knowledge, or are driven by a real hunch as to the correct answer. In other words, the prevailing practice in-

[15] Wood, E. P., "Improving the Validity of Collegiate Achievement Tests," *Journal of Educational Psychology*, Vol. 18 : 18-25, 1927.

dicates that they should be instructed "not to guess" on such exercises but their responses should be treated as though they did guess.

Correcting for chance or guessing is another point about which there has been and still is considerable controversy. Many investigators have worked on this problem and much still needs to be done. The available evidence relative to the effects of "correction for chance" upon reliability is still debatable. Most of the studies indicate a slight lowering of the reliability, although the study by Ruch and DeGraff,[16] the most extensive to date, points toward a slight increase in reliability when the correction is applied. The evidence is quite consistent in showing that the correction for chance increases the validity of test scores, especially in true-false or Yes-No, two-response and three-response exercises.

The common method of correcting for chance is to deduct the number of exercises answered incorrectly from the number answered correctly (Right—Wrongs). For example, a pupil in taking a true-false test attempts 76 items, answering 62 exercises correctly and 14 incorrectly. His corrected score is 62 less 14, or 48. Scores less than zero are not ordinarily assigned in tests of this type. Thus, a pupil answering correctly only 24 exercises on the same test, while making 40 errors, would receive a corrected score of zero. The correction for chance can be generalized into a simple formula which takes care of all types of alternate- or multiple-response tests without regard to the number of possible choices involved. The corrected score is equal to the number of exercises answered correctly, minus the number answered incorrectly, divided by the number of possible choices less one. In symbols this formula is as follows:

$$S = R - \frac{W}{(N - 1)}$$

in which

R is the number of exercises answered correctly,
W is the number answered incorrectly, and
N is the number of choices.

[16] Ruch, G. M., and DeGraff, M. H., "Corrections for Chance and 'Guess' *vs.* 'Do not Guess' Instructions in Multiple-Response Tests," *Journal of Educational Psychology*, Vol. 17, pp. 368-375, 1926.

It may be well, in concluding this discussion on chance factors in objective tests to point out that teachers who do not wish to concern themselves about correcting test scores for chance may avoid this problem by simply increasing the length of the objective test. The addition of 10 or 15% more items than was originally planned will largely eliminate the need for making corrections.

Discussion Exercises

1. By way of review, what are the chief causes for the dissatisfaction of teachers with the traditional or essay-type of examination?
2. What are the fundamental differences between standardized tests and informal objective tests?
3. What reasons can you advance for the general conclusion that there is little conflict between informal objective tests and standardized tests?
4. Enumerate the advantages of the objective test over the traditional essay type of examination.
5. Show how the factor of sampling alone influences accuracy of measurement.
6. What are the most serious disadvantages of the objective examination?
7. State in your own words if possible as many as six of the principles for use in constructing objective exercises.
8. What are the main factors determining the length (number of items) of an objective examination?
9. What reasons can you advance for the prompt revision of a new objective examination before it is used again? What types of data may be used in such a revision?
10. Why is it important that examinations or objective tests should not be so difficult as to cause 100% failure or so easy as to permit 100% success?
11. Why should items in alternate-response type tests be arranged so that the order of answers is in a random order?
12. Suggest arguments pro and con on the matter of using different types of objective exercises in the same examination.
13. To what definite sources may the teacher turn in the selection of objective test materials?
14. In what specific ways do objective tests greatly simplify the teacher's problems of measuring and grading pupil achievement?

15. Is the teacher's judgment of pupil accomplishment entirely eliminated with the use of objective tests? If not, at what points does it remain?
16. Point out the distinguishing features of the two major classes of objective testing techniques.
17. To what extent does chance in the form of guessing actually enter into the results in the different types of objective exercises?
18. Show how the general formula for correcting for guessing is applied in the case of an alternate-response test, a three-response test, and a five-response test.

Selected References

Banker, H. J., "Significance of Teachers' Marks," *Journal of Educational Research*, Vol. 16 : 159-171 and 271-284, 1927.

Buckingham, B. R., *Research for Teachers*. New York : Silver, Burdett and Company, 1926.

Foster, R. R., and Ruch, G. M., "On Corrections for Chance in Multiple-Response Tests," *Journal of Educational Psychology*, Vol. 18 : 48-51, 1927.

Greene, H. A., "A New Correction for Chance in Examinations of the Alternate-Response Type," *Journal of Educational Research*, Vol. 17 : 102-107, February, 1928.

Hahn, H. H., "A Criticism of Tests Requiring Alternative Responses," *Journal of Educational Research*, Vol. 6 : 235-40, 1922.

Holzinger, K., "On Scoring Multiple Response Tests," *Journal of Educational Psychology*, Vol. 15, pp. 445-447, 1924.

Lang, Albert R., *Modern Methods in Written Examinations*. Boston : Houghton Mifflin Company, 1930.

McClusky, H. J., and Curtis, F. D., "A Modified Form of the True-False Test," *Journal of Educational Research*, Vol. 14 : 213-224, 1926.

Odell, C. W., *Traditional Examinations and New-Type Tests*. New York : The Century Company, 1928, 400 pp.

Orleans, J. S., and Sealy, G. A., *Objective Tests*. Yonkers, N. Y. : World Book Company, 1928.

Paterson and Langlie, "Empirical Data on the Scoring of True-False Tests," *Journal of Applied Psychology*, Vol. 9 : 339-48, 1925.

Paterson, D. G., *Preparation and Use of New-Type Examinations*. Yonkers, New York: World Book Company, 1925, pp. 70-71.

Paterson, D. G., "Do New and Old Type Examinations Measure Different Functions?" *School and Society*, Vol. 24 : 246-248, 1926.

Ruch, G. M., *The Objective or New-Type Examination.* Scott, Foresman and Company, 1929.

Ruch, G. M., and DeGraff, M. H., "Corrections for Chance and 'Guess' *vs.* 'Do Not Guess' Instructions in Multiple-Response Tests." *Journal of Educational Psychology*, Vol. 17 : 368-375, 1926.

Ruch, G. M., et al., *An Investigation of Objective Examinations in the Social Studies.* Chicago : Scott, Foresman and Company, 1926.

Ruch, G. M., and Stoddard, G. D., *Tests and Measurements in High School Instruction.* Yonkers, New York : World Book Company, 1927.

Russell, Charles, *Classroom Tests.* Boston : Ginn and Company, 1926.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York : Silver, Burdett and Company, 1928.

Symonds, P. M., *Measurement in Secondary Education.* New York : The Macmillan Company, 1927.

Weidemann, C. C., *How to Construct the True-False Examination*, Teachers College, Columbia University, Contributions to Education, No. 225, 1926.

Wilson, G. M., and Hoke, K. J., *How to Measure.* New York : The Macmillan Company, 1928.

Wood, Ben D., *Measurement in Higher Education.* Yonkers, New York : World Book Company, 1923.

CHAPTER V

THE USES OF STANDARDIZED EDUCATIONAL TESTS

This chapter presents a discussion of the following uses of standardized educational achievement and mental tests as instructional and supervisory instruments:

a. The comparison of the achievement of pupils, classes, and schools with each other and with norms and standards.
b. Determining the efficiency of instruction.
c. Class and individual diagnosis by means of achievement tests.
d. Classification, gradation, and pupil guidance by means of achievement test results.
e. Classroom uses of mental tests.

The problems of improving the procedures and devices by which classroom achievement may be measured have been emphasized in the two chapters immediately preceding. In contrast with the emphasis in these chapters, the problems of educational measurement are considered in this chapter from the standpoint of the standardized educational test as an instructional and supervisory instrument.

48. Tests in Relation to Classroom Instruction. So much comment has been made about the value of standardized educational tests as supervisory instruments, as aids to school surveys, as research equipment, and as tools for administrative purposes, that very often the classroom teacher overlooks their real value in the solution of his own instructional problems. Yet this is where the most vital and important uses of such tests are to be found. The development of reliable, valid, and highly detailed measuring instruments has caused him to modify his previous conceptions of the uses of standard tests. Earlier experience with the more formal types of educational tests led the teacher to feel that tests were merely time-consuming devices used for checking up

on his teaching efficiency, and from which he at no time received any constructive help in the improvement of his instruction. The more modern conception of standard tests is quite in contrast with this notion. It implies their continuous use as instruction progresses. Upon the completion of an instructional unit the teacher calls to his aid a section of a detailed diagnostic or analytical test as a means of determining for himself whether his emphasis has been so placed that the desired habits or skills are properly developed in the pupils. In a sense this means a continuous testing program, which experience with the other conception of the use of tests indicates is the only way that standard tests will ever come to function at their highest efficiency as instructional instruments in the classroom.

I. Instructional Uses of Achievement Tests

49. **For Purposes of Comparison.** Since the earliest beginnings of group instruction, classroom teachers have wished to know just how their pupils have compared in attainment with other similar pupils and classes. Until after standard tests were developed it was practically impossible to secure this information. Now the giving of a standard test in arithmetic, spelling, reading, or other school subject makes fairly easy a comparison of the results from a class with the norms and standards established for that subject and grade. For example, the use of the *Haggerty Reading Examinations, Sigma 3* by a teacher furnishes the following data for comparison:

1. Individual pupil reading scores
2. Individual pupil reading ages
3. Individual pupil reading quotients
4. Average or typical reading scores for the class
5. Average reading ages for the class
6. Average reading quotients for the class

In this way a teacher may discover what should be expected of a class as well as how the class is progressing in a particular school subject.

Comparisons with other classes within the system in which the teacher is working, within the same building and even

with different sections of classes in charge of the same or different teachers can be made on a basis of objective standards that have been derived for the various tests. Another sort of comparison which is often useful is that between the attainment of a class at the beginning and the end of a semester's or year's work, or at shorter intervals in the course of a semester. Each of these comparisons has its own peculiar value in assisting the teacher to determine the relative attainment and progress of his class at a given time. Such comparisons also serve to stimulate the individual pupil to greater effort and place upon him a more definite feeling of responsibility for his own development.

50. **For Measuring the Efficiency of Learning.** Such general comparisons as are cited above are of great value in themselves, by way of indicating whether classes are up to standard in their achievements. Equally important is the determination by means of comparative studies whether the act of teaching itself may not be improved. Ambitious teachers everywhere are looking for the best methods of instruction in their fields. Teaching methods, which in the last analysis can be evaluated only in the classroom by the classroom teacher, will be most effectively evaluated by means of standardized tests. Instructional units within the course of study must also be evaluated. The measurement of the effect of certain types of drill exercises, the determination of the specific strengths or weaknesses of groups or classes constitute uncounted opportunities for the use of these valuable instructional devices. Only through the use of large quantities of such available material in connection with numerous such simple yet significant investigations by the classroom teacher will the multitude of now unanswerable questions in this field be answered.

51. **For Purposes of Class Analysis and Diagnosis.** Very often a teacher, at the beginning of a school term, wishes to obtain advance information concerning the proficiency of his classes in certain subjects, and their general preparation for the work. It is essential for him to know their weaknesses and their strengths, in order to direct their work in such a manner that the best results will be obtained. He needs to know the background his pupils have been given for the

work they will be expected to master during the ensuing year. Many standard tests are now available which permit this type of use. It is not always necessary for a teacher to employ a special diagnostic test to secure the required general data on the relative abilities of the class. The ability to read silently is the basis of proficiency in so many subjects, that he should certainly secure a picture of the reading ability of the class. The results should indicate whether the class as a whole or the individual members of the class are able to interpret the printed page with facility, and so carry on their work without great assistance. It is also possible to test in a like manner for reasoning, judgment, and for other general qualities, as well as for knowledge of specific subject matter.

Not only is this preliminary diagnosis of great value to the teacher, but it has also been found desirable and valuable to check progress or advancement from time to time by means of objective tests. Tests of achievement will reveal whether the class as a group is moving together, or whether there are more or less well-defined sub-groups which seem to need special attention. Frequently such conditions furnish a justifiable basis for dividing a grade or class into sections for such corrective treatment. Where classes are divided into several sections as is often done in larger schools, many competent educators feel that the pupils should be arranged so that groups of approximately equal ability are placed together. The objective test is the best means for making this adjustment so that pupils can move forward in groups of nearly equal power to learn and ability to achieve.

An excellent illustration of this need is given in Table 3 showing the rather startling range of mental ability found in a typical 9th grade class. A teacher confronted with a class ranging in ability from above twelfth grade ability to below seventh grade is faced by a hopeless task if he is expected to bring all members of this class up to the same level of proficiency. Particularly is this true when the range of ability is wholly unsuspected or measured by guess rather than by reliable tests as it is in so many cases. The systematic use of tests for class diagnosis constitutes a real source

of professional protection to the classroom teacher and his pupils.

Table 3

DISTRIBUTION OF MENTAL ABILITY OF A 9TH GRADE CLASS

(According to Grade Location of Ability on Terman Group Test)

Grade Location	Number of Pupils
Above 12	6
12	4
11	12
10	8
9	19
8	8
7	6
Below 7	4
Total Number of Cases	67

52. **For Purposes of Individual Pupil Diagnosis.** Closely connected with the use of the test for the analysis and diagnosis of a class or grade is its use for the determination of the difficulties and variabilities of each individual pupil. While, in general, individual differences are not so marked as to preclude reasonably efficient class instruction yet the more that is known about each child's weaknesses and strengths, the greater are the possibilities for success on the part of the teacher instructing the group. The test results should be studied especially in the light of each pupil's individual attainments, and points of difficulty. The critical analysis of each pupil's test score may very likely be a means of clearing up wholly unsuspected troubles which would otherwise continue to hamper the child and to reduce his chances for proper advancement. While this type of individual diagnosis has great possibilities, it becomes valuable only when it is definitely tied up with remedial material so devised that each child may be aided in correcting his own weaknesses.

Many illustrations of this type of material will be found in later chapters in this book dealing with diagnosis and remedy in the special subject fields.

It must be clear from the foregoing discussion that one of the main purposes in the use of tests is achieved when precise and accurate diagnosis of individual pupil weakness is followed promptly by valid remedial drill.

53. **For Pupil Gradation.** Teachers and supervisors find the problem of pupil placement one of the most difficult situations which they have to face. The indefinite lines of division between the grades and the wide overlapping of ability within the grades reveal the fact that our technique of pupil classification has been extremely crude. Practically all investigations have shown the typical grade assignment of pupils to be inaccurate and unsatisfactory. This could scarcely be otherwise in view of the methods commonly used. Chronological age and physical size are the two pupil characteristics used most frequently even under conditions where much more exact information might be readily secured. The proper grade placement of pupils implies that so far as possible individuals who are normal for their group will be placed together. This means that pupils who are approximately alike in their chronological age, educational achievement, physiological, mental, social, and moral development should ordinarily be placed together for instructional purposes. Not all of these qualities lend themselves readily to objective measurement, but a number of them do, and within these limits the results of objective measurements should be used in determining the pupil's placement in his group. Through the development of reliable grade and age norms, based upon the achievement of groups of children on standardized tests, a valuable instrument for the establishment of much more exact grade lines is made available.

By means of a simple procedure, results from several standardized tests, even though they may express their results in different units, can be combined into a simple graphic record and used as a valuable aid in pupil classification. The technique is simple and valuable regardless of whether a complete reclassification of the school or grade is planned

or merely the proper placement of a few new pupils entering the system for the first time. The procedure by which test scores may be graphically compared and used for classification purposes is explained more completely in Section 127 of Chapter IX.

54. **For Pupil Guidance.** The schools are constantly being subjected to criticism because of apparent waste which arises from their frequent failure to identify the special abilities of their pupils and challenge these children to greater efforts. This is one aspect of educational guidance. It is also charged that little or no guidance is given students away from fields in which they apparently have no aptitude. With the objective devices now available for the measurement of general as well as specific abilities of children neither of the situations need exist. Teachers, principals, and college administrators have found that test records obtained early in the institution's contact with the student prove to be extremely valuable aids in handling disciplinary cases and helping students to adjust themselves in many other ways. The majority of such administrative problems as disciplinary cases in schools or colleges arise through the failure of the school system properly to stimulate and occupy the student's mind. Many of the reasons for such difficulties may be made clear to the teacher by the wise use of properly selected tests. The necessary adjustments can then be made and correct a situation which need not exist if properly handled.

II. Uses of Mental Tests

55. **Mental Tests in the Classroom.** The meaning of mental or intelligence tests in general has been discussed very briefly in a preceding chapter (see pages 20-21). It remains to consider intelligence tests from the standpoint of their practical use to the classroom teacher. A large part of the rapidly growing popularity of intelligence tests among teachers and supervisors may be traced to three main causes; namely, (1) the tests themselves have been greatly improved in the accuracy and value of the resulting measures; (2) a larger proportion of school officers have become

more intimately acquainted with intelligence tests and testing procedure with a correspondingly greater appreciation of the functions they serve; and (3) the changes in modern conceptions of education and attitudes toward it have made the utilization of such devices almost imperative. The stimulus given mental testing as a result of the gigantic army testing program during the World War resulted in the production of many types and forms of individual and group intelligence and general ability tests. Some of these have stood the wear and tear of time and some have been superseded by other more carefully constructed or more modern devices.

The need for information such as is provided by any competent mental test is so great that no classroom teacher can be assured that her procedure with a given pupil is sound unless she has access to such exact information. The majority of group intelligence examinations must be used carefully and interpreted cautiously. Most of such group tests measure with sufficient accuracy for group comparisons but thus far the results from even the best of the group intelligence tests should not be taken too seriously when it comes to individual comparisons. Many group mental tests permit the determination of mental ages (M.A.) and I.Q.'s (Intelligence Quotients) from the test scores. These derived scores if based on group test results are undoubtedly too unreliable for use where close decisions are involved, but are, in the main, sufficiently accurate for most school situations. If it is important that the mental age or the I.Q. of an individual pupil be obtained and only group tests can be used, not less than two forms of the test should be given at different periods. Even this procedure may result in conclusions which in many cases may be seriously erroneous. However, it need hardly be pointed out that a measure of intelligence based on a reasonably reliable group test is better than anyone's subjective estimate of the ability of the subject.

The many important problems which arise in connection with the administration and interpretation of mental tests in the classroom are treated in detail in Chapter XI, *Mental Tests as Diagnostic Aids.*

Discussion Exercises

1. From the standpoint of the classroom teacher what are the major values of educational tests?
2. Enumerate the principal uses of tests for comparative purposes.
3. Show how the standardized test reveals evidence on the efficiency of instruction.
4. Suggest a procedure by which properly designed tests may be used for class diagnosis.
5. What does the evidence tend to show with regard to the actual range of ability which a teacher may encounter in a given class?
6. Show how a properly designed test or series of tests may be used for the remedy of defects of individual pupils.
7. What features of standardized tests make them particularly useful for pupil gradation and guidance?
8. Prepare an individual pupil record card which would be useful as a basis for pupil diagnosis.
9. Why is the practice of publishing (giving publicity to) the results of intelligence tests an extremely dangerous procedure?
10. Show how the interpretation of achievement test results is greatly simplified by having mental test results available at the same time.

Selected References

Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching.* Boston : Houghton Mifflin Company, 1931.

Courtis, S. A., *Why Children Succeed.* Detroit : Courtis Standard Tests, 1925.

Kelley, T. L., *Interpretation of Educational Measurements.* Yonkers, New York : World Book Company, 1927.

Lang, A. R., *Modern Methods in Written Examinations.* Boston : Houghton Mifflin Company, 1930.

Madsen, I. N., *Educational Measurement in the Elementary Grades.* Yonkers, New York : World Book Company, 1930.

Odell, C. W., *Educational Measurement in High School.* New York : The Century Company, 1930.

Ruch, G. M., Stoddard, G. D., *Tests and Measurements in High School Instruction.* Yonkers, New York : World Book Company, 1927.

Smith, H. L., Wright, W. W., *Tests and Measurements.* New York : Silver, Burdett and Company, 1928.

Webb, L. W., Shotwell, Anna M., *Standard Tests in the Elementary School.* New York : Ray Long and Richard Smith, Inc., 1932.

Wilson, G. M., Hoke, K. J., *How to Measure* (Revised). New York : The Macmillan Company, 1928.

Woody, Clifford, Sangren, P. V., *Administration of the Testing Program.* Yonkers, New York : World Book Company, 1933.

CHAPTER VI

USING STANDARD TESTS IN THE CLASSROOM

This chapter deals with the following points involved in the effective use of standard tests in the classroom:

a. The teacher's responsibility for the use of standard tests.
b. Factors determining when tests should be given.
c. Teacher's responsibility for the use and interpretation of test results.
d. Specific steps in the inauguration of a testing program.
e. The teacher's problems as a basis for the testing program.
f. Criteria for the selection of tests.
g. Suggestions for conducting a testing program.

The value of the educational test to the classroom teacher is directly proportional to the extent that the results obtained are translated directly into improved teaching practices for the teacher and learning conditions for the pupil. The problems involved in securing these results in the most effective and economical manner are the problems of the testing program.

I. Giving the Tests

56. The Teacher's Responsibility. In the earlier stages of development of standard tests it was believed that the most valuable results came from their use in a periodical survey by persons other than the classroom teacher. Recently it has come to be more generally accepted that as many of the tests as possible should be given by the classroom teacher. This seems to be especially true in the case of tests which furnish information of special importance in the improvement of classroom instruction. In addition to allowing the classroom teacher to become acquainted with the technique

of testing it gives him a first hand opportunity to observe the reactions of his pupils in the various test situations. On this account it is believed that wherever test results are to be used definitely as a basis for the discovery of individual pupil difficulties they should as far as possible be administered and interpreted by the classroom teacher himself. However, where the test results are used for a survey of achievement in the entire school or system, it is less important for the teacher to have an intimate contact with the testing program. As a matter of fact many school administrators prefer not to have the teachers give the tests when they are used for such survey purposes. Furthermore in the case of certain types of tests such as the individual mental examinations which require special training for their administration, it is probably wise to discourage the average classroom teacher from attempting their use.

57. **Scoring and Interpreting Tests.** The matter of the responsibility for scoring and interpreting the test results constantly arises as an administrative problem in the smaller school. Teachers are likely to feel that the responsibility for scoring standard tests given for supervisory purposes should not fall to them. A part of this difficulty arises through a failure on the part of the administrators to make perfectly clear to the members of the teaching staff their responsibility with regard to this type of work at the beginning of their term of service. Most teachers, if given a suitable amount of time, do not seriously object to scoring test papers for their classes, particularly when they come to realize that this work is almost certain to reveal information which will be extremely significant to them in the improvement of their teaching practice. Teachers should consider it an opportunity rather than an additional responsibility to correct the test papers for their classes, and to study the results revealed by the tests. If a real interest in the outcome of the testing program is stimulated by the supervisory officers there will be little difficulty in inducing the teachers to help in the scoring and interpreting of the test papers from their classes.

II. When to Give the Tests

58. Periodical vs. Continuous Use of Tests. Tests may be used periodically or they may be used somewhat constantly in the classroom. The type of testing followed depends somewhat upon the purposes the tests are to serve and the nature of the tests selected. If the tests are the common survey tests of general achievement they are usually given early in the school term, and then again a few days before the end of the school term. This procedure permits the teacher to determine the improvement which his pupils have made during this period. If he is a wise teacher he will realize that a portion of this measured improvement is due to sheer increase in maturity on the part of his pupils and that only a portion of it is the result of his instruction. Tests which are used definitely for survey purposes, and are given only once during the school year, are frequently administered at or near the end of the year. This is probably one of the least important times for tests to be given, since the entire school year is gone and there is no opportunity for the teacher to attempt to do anything about the conditions revealed by the tests. If only a single cross-section of the school is taken this should undoubtedly come early enough in the school year to permit the teachers to profit from the findings. This periodical use of educational tests to measure class or individual pupil improvement is by far the most prevalent practice.

A further refinement of the idea of using tests early in the school year is found in their use immediately following the completion of the instruction on a particular unit of subject matter. This means that many tests each measuring a rather narrow field must be available. This, of course, is exactly what is being done in the development of the detailed diagnostic tests which are now proving so useful. Unit achievement tests, each designed to measure a specific area of the course, are also proving to be popular with both teachers and pupils. By using these narrow-function tests immediately after the completion of the teaching of a specific subject-matter unit the teacher secures immediate information about the weaknesses of his class. The special

inadequacies of his instruction are thus made clear, and he can proceed at once to set up a remedial program before the class has moved on to other subject matter. This procedure represents the continuous use of tests as the basis for remedial work. The same type of information provided by the use of these numerous narrow-function tests which guides the teacher in setting up the remedial program is also useful in organizing instruction in the future to prevent the appearance of these recognized weaknesses.

III. The Basis for the Testing Program

59. **Clear-cut Teaching Problem as Basis.** One of the most common errors made by teachers and supervisors is the inauguration of a testing program without the formulation of a clear-cut problem, the solution of which can be most advantageously reached through the use of standard tests. The more clearly defined this problem is the better, for the testing program will thus be more limited in extent and more intensive. The best constructive supervisory work will result from careful intensive cultivation of a limited field. If the work is undertaken in this way much time will be saved and one of the most common criticisms avoided; namely, that the time of the pupil and teacher is taken for the testing and nothing ever comes of it. The teacher and pupils have a right to profit from a knowledge of the conditions revealed by the testing.

60. **Steps in a Testing Program.** A tentative outline for a testing program is presented here as a suggestion for the general organization of the work.

1. Formulate and state a clear-cut teaching or instructional problem in the solution of which data from standardized tests will prove valuable.
2. Select the best available tests for the purpose.
3. Determine the time and method of administration of the program.
4. Train the staff to be employed in the administration of the program so that uniformity of conditions may be guaranteed.
5. Administer the tests.
6. Score and tabulate the results.
7. Analyze and interpret the results.
8. Propose a program which will result in the elimination or improvement of the conditions revealed.

A further discussion of these steps and their elaboration in connection with the problem of administering a testing program constitutes the purpose of the next pages.

61. **Illustrations of Problems.** Problems suitable to form the basis for a testing program are to be found in almost all of the fields of education. Frequently these problems overlap. It is not uncommon for one test or a series of closely allied tests to contribute to the solution of several problems. The problems listed below are classified in accordance with their interest to administrators, supervisors, and teachers. It is clear, of course, that while the list is suggestive, it is not exhaustive.

A. Problems Primarily of Interest to Teachers and Supervisors

1. The determination of how near the pupils and the classes are to the standards in the different subjects.
2. The determination of the progress of the class in the different school subjects over a given period of time.
3. The determination of whether different subjects and phases of subjects are being properly or unduly stressed, as indicated by relative accomplishment.
4. The discovery and diagnosis of defects of individual pupils in the various subjects or in particular phases of a subject as a basis for a remedial program.
5. The determination of whether the pupil or the class is working up to capacity. Pupils of greater mental capacity should be expected to accomplish the most in their school work.

B. Problems Primarily of Interest to Administrative and Supervisory Officers

1. The division of classes into two or more sections according to ability. Either intelligence or educational achievement tests or both may be used for this purpose.
2. The selection of pupils for special classes such as classes for the exceptionally bright or exceptionally dull pupils or for pupils having special defects in certain subjects.
3. The determination of the efficiency of the school as a whole by comparison of obtained scores with standard scores and with scores made by other schools or grades.
4. The determination of whether the proper emphasis is given to all subjects or whether some subjects are overstressed.
5. The comparison of different methods of instruction or comparison of new methods with the methods of instruction.

6. The determination of the general intelligence or achievement level of a grade, a school, or of a system.
7. The measurement of the progress of a grade, a school, a system, for a semester, a year, or for any given period of time.
8. The determination of whether or not the grade, the school, or the system is achieving what can be fairly expected for achievement.
9. The determination of the misplacement of pupils in grades or sections.
10. The proper grade classification of all new pupils entering the school system.

IV. The Criteria for Selecting Tests

The second step in planning a testing program is the selection of the tests to be used. In this connection there are certain criteria which the teacher or supervisor will do well to keep in mind.

62. Validity. Is the test *valid?* Does it really measure the qualities, abilities, and skills, that it is supposed to measure? For a test to possess validity it is necessary that the materials actually included in the test be of prime educational importance, that the questions sample widely among the essentials over which complete mastery on the part of the pupils can be reasonably expected, and which they are given an opportunity to master in the course of study.

Validity may be defined more specifically as an expression of the adequacy with which a test measures the qualities, abilities, skills, and information it is designed and supposed to measure. In a general way it may be expressed in terms of the extent to which the test sets up a situation calling into play the use of the skills, or abilities which experienced observers consider fundamental to success in a given field. Such judgments are represented by the opinions of authors of textbooks and courses of study, experienced teachers, and recommendations of committees, and other qualified authorities. Numerous other validation techniques are suggested in Chapter VII. Validity may be expressed statistically in terms of the degree of relationship (correlation) found between the particular measuring device and some other similar device, the usefulness of which has previously been established. An illustration of this type of validation is given in

Problem 39 of the *Work-Book in Educational Measurements.*[1] In this case the criterion used is the judgment of the teacher who has had extensive contact with the individual pupils.

The source from which material in a test is selected is one index to the validity of the test. For example, a spelling test composed of words selected from an unabridged dictionary by taking the first word found on each tenth page of the book would certainly not be a valid test for at least two reasons. In the first place, the words would be of unknown social usefulness. In many cases it would not matter whether the pupils could spell the words or not. In the second place, the test would be lacking in validity because the pupils taking it would doubtless have had no opportunity to learn to spell these words.

The validity of a test may be thought of as being *general,* or it may be considered as being *specific.* Many tests are valid in a general sense but are lacking in validity in the specific sense. For instance, a survey test designed to measure general achievement in a particular subject must be validated in terms of its content of the basic items in the course as it is taught, not in a single class in which definite points of view and certain facts have been emphasized, but in many different schools. A teacher may easily construct a test which will have much greater specific validity for his purposes and his point of view than any commercial standardized test could possibly have. This point has already received considerable attention in connection with the discussion of validity of content of informal objective examinations in Chapter IV.

A useful illustration of the application of the method of validating test material in terms of judgment of qualified authorities is found in the development of the *Iowa Silent Reading Test, Advanced Examination.*[2] The following outline shows the unit skills contributing to the high school pupil's ability to read and to work with books, which are measured by these reading tests.

[1] Greene, H. A., *Work-Book in Educational Measurements,* 1928.

[2] Greene, H. A., Jorgensen, A. N., and Kelley, V. H., *Iowa Silent Reading Test, Revised, Advanced Examination.* Yonkers, New York : World Book Company, 1931.

Test 1. Paragraph Meaning
- A. Science
- B. Literature

Test 2. Word Meaning
- A. Social Science
- B. Science
- C. Mathematics
- D. English

Test 3. Paragraph Organization
- A. Selection of Central Idea
- B. Outlining

Test 4. Sentence Meaning

Test 5. Location of Information
- A. Use of Index
- B. Selection of Key Words

Test 6. Rate of Silent Reading

Compare with this list of measurable skills the major objectives of reading instruction given in Chapter XIV, pages 352 to 355. The case for the validity of this reading examination rests on the exactness with which these objectives of reading have been paralleled in the test parts comprising it. In this case the validity of the test is determined by the extent to which it measures the desirable skills in silent reading as recognized by specialists in this field. In the case of achievement tests based more generally upon information rather than skills, the validity of the test depends more largely upon the opportunity the pupil has had to master the content measured in the test. In such a situation the teacher himself is probably the best judge of the validity of the test, since he knows best what material he has taught the class.

The general concept of validity may be made clear if the reader will think of the test situation as a small section (sampling) of a larger situation in life. The more nearly the test situation reflects the actual situation in life the more valid it becomes as a measure. In the attempt to devise tests which would be valid for the selection of persons for different types of activities during the War the last and final test was an exact duplicate of a life situation. Truck drivers who were being tested for their ability to take a supply or hospital truck into a shell-ridden area under cover of darkness were tested by actually having them drive a truck over

shell-wrecked roads at night without lights and under fire. In such a case the test situation was a life situation. The successful passing of such a test was proof of ability. Inferences as to ability were not necessary. The validity of the test was practically perfect. The direct application of this conception of validity to standardized or informal tests is, under practical circumstances, often hard to approach but where it is possible it is a useful criterion.

There is a growing tendency among users of standardized tests to place the responsibility on the makers of the tests for the establishment of the facts concerning the validity and reliability of the test. In the examiner's manuals for the better types of standardized tests complete statements of the validity and the reliability of the tests will be found. Persons interested in the critical examination of a test soon learn to look for these facts first.

63. **Reliability.** A test is said to be reliable when it functions consistently. Consistent performance in a test arises from two sources, (1) the adequacy of the sampling represented by the test, and (2) the variation in the human response itself which has nothing to do with the content of the test. These points are discussed in order in this section. The first of these factors in reliability may be controlled somewhat by selecting the test items carefully and extensively from the field which it is supposed to measure. This is illustrated clearly in the case of the examination of carloads of grain for the market. It is obviously impossible for the purchaser to examine minutely every bushel of corn or grain in the car. It is possible, however, for specimens of the grain to be taken from different places in the car and from different cars. This process is called sampling. On the basis of these samples the car of grain is graded as to quality and on the basis of this quality its price is determined.

In a similar manner exercises or items of information are sampled from the possible items suitable for a test in a given field. The accompanying diagram (Figure 3) illustrates the effect of sampling on the reliability of test exercises based on a certain limited field of information. Each of the 40 rectangular spaces in Figure 3 represents an item of informa-

Pupil A

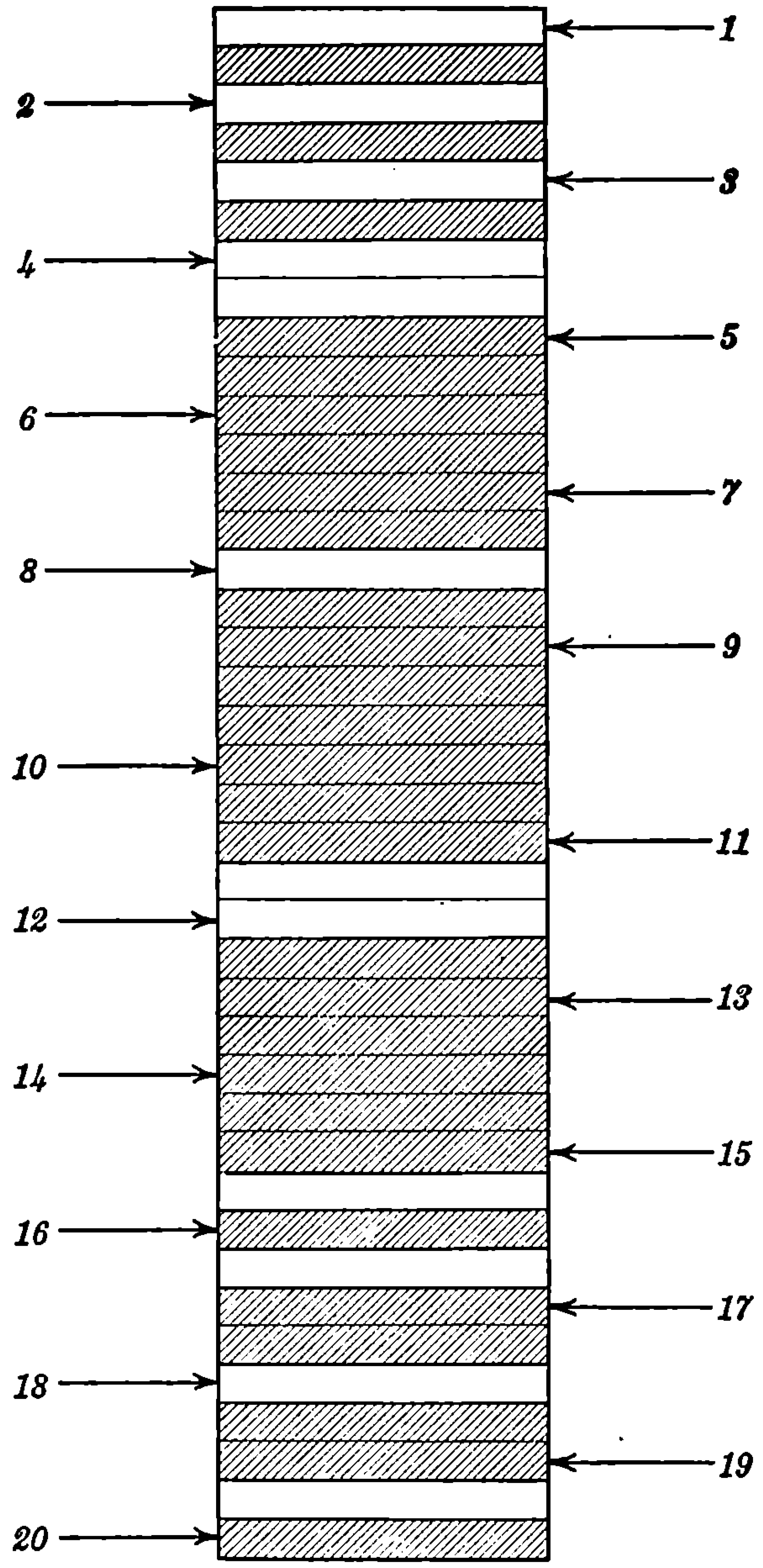


FIGURE 3. THE PRINCIPLE OF SAMPLING

tion which an individual pupil (Pupil A) has had a chance to learn. The 28 shaded spaces represent the 70% of the items which he has actually mastered. The 12 unshaded portions are items which he has not learned. These items are thus 30% of the total. Let us assume that a very limited sampling at one end of the field of his information is made. If Exercises (items) 1 to 5 inclusive are selected he will fail on all but one, item 5. On the other hand, if

five exercises from the opposite end of the field of information are chosen he will fail on only one, item 18; thus a variation of from 20% to 80% on the items. From this it is clear that it makes a difference where the sampling is taken. This factor is largely taken care of, however, if the sampling is adequate in extent. If the even numbered items are selected the pupil will fail on half of them. If the odd numbered exercises are chosen he will fail on only two. This results in a variation in his score of from 50% to 80% due largely to chance. If the twenty items are selected the pupil fails on 35% of the exercises, and answers 65% correctly. It is to be noted that as the number of items selected is increased the pupil's success on the test more nearly approaches the actual amount of his information in this field. Thus it is clear that the extent of the sampling which the exercises in a test represent is an important factor in its reliability.

Increasing the reliability of tests means that chance factors such as temporary physical disturbances, fatigue, etc., which affect the consistency with which individuals react to tests must be more largely eliminated from the results by increasing the length of both testing periods and tests. Serious doubts are beginning to be cast upon many general survey tests which limit their period of sampling to one period of two, five, or even ten minutes. Extremely limited and narrow skills may be sampled in that time but it is very doubtful if it is possible to sample in that period of time such highly complex skills as many of our school subjects represent. However, the amount of time required for giving and scoring a test should be considered with reference to the amount of information yielded. The time and money costs of tests may come to be calculated in terms of the cost per unit of information secured from their use.

The reliability of a standardized test should be stated in the examiner's manual. The coefficient of reliability is a statistical expression of the consistency of performance of the test and is indicative of the reliance which may be placed on scores obtained from its use. Reliability coefficients are indicated as decimal fractions of one, usually ranging in value from .50 for tests of low reliability to .95 for tests of high

reliability. In general, in order to be acceptable for classroom use a test should have a reliability of .88 to .92 when used in a single grade.

Reliability, which is the index to the consistency of performance of a test, is commonly expressed in terms of the degree of correspondence between results from consecutive trials of the test, or between results from two forms of the same test. This procedure is illustrated concretely in the case of Problem 14 on Page 200 of Chapter VIII and Problems 40 and 41 of the *Work-Book in Educational Measurements.* Since the process is somewhat technical and involves methods of correlation which have not been explained up to this point in this book, the student should probably postpone the examination of these illustrations until the statistical methods involved have been explained and mastered.

64. Objectivity. A test is *objective* when it can be administered and scored without the results being affected by the teacher's personal judgment. Historically, objectivity is one of the first qualities of examinations to be noted by teachers. The need for the removal of the subjective factors in the marking of examination papers is one of the reasons back of the development of standardized tests. In fact, it is mainly the removal of the personal judgment factors by the objective form in which the exercises are stated, and the use of mechanical methods of scoring the exercises, that standardized tests are now considered to be superior to the traditional type of examination. Objectivity is an important requirement of a good test, but must be considered as secondary to validity and reliability, although it may contribute to each of these qualities indirectly. *Objectivity in a test exercise makes for the elimination of the opinion of the person who scores it.* This means greater accuracy in the grading of the exercises, and naturally greater reliability of measurement results.

In general, objective test exercises are stated in such a manner that only one answer satisfies the requirements of the statement. A common form of objective exercise is the recall type. Other types are the true-false, multiple response, matching exercises, etc. These and other forms of

exercises are illustrated on pages 81 to 84 of this volume. The distinct advantage of selecting perfectly objective exercises for use in standardized educational tests is that there can be a perfect agreement as to what is the correct answer.

TABLE 4

MARKS ASSIGNED BY TEN TEACHERS TO AN ESSAY-TYPE EXAMINATION IN CIVICS, AND TRUE-FALSE EXAMINATION OVER THE SAME MATERIAL

Teacher	Mark Assigned	
	Essay-type (10 Questions)	True-false (40 Exercises)
A	84	31
B	80	32
C	88	31
D	76	30
E	82	31
F	90	31
G	85	31
H	82	31
I	83	31
J	80	31
Average.......	83	31
Average error (deviation from the average) ..	3	.2

This means that outside of purely chance errors there should be no variation in the scores assigned to a given test paper by different persons.

The effect of objective exercises on the accuracy of marks assigned to a test or examination paper is illustrated in Table 4. This table shows the relative accuracy of marks assigned by a group of ten teachers to two types of test papers written

by the same pupil on the same subject matter. Each teacher graded each paper independently and at different periods so that there was little chance for memory of the grade assigned previously to carry over. One of the papers was a typical ten question essay-type examination in the field of Civics. The other paper was a forty-exercise true-false test based upon the same subject matter. It will be noted from Table 4, that the average mark assigned was 83, on the essay-type examination paper, and 31 points on the true-false objective test. It will also be noted that the highest mark assigned to the essay-type paper was 90 and the lowest was 76 points. All but two of the teachers agreed, however, on a mark of 31 as correct for the true-false examination. The relative objectivity of the two types of examinations is shown definitely by the variation in the marks assigned by these ten teachers. In the essay-type examination each teacher missed the average score by an average error of three points. In the case of the objective examination paper the average error was .2 point, and this error was contributed by two individuals, Teacher B and Teacher D.

Objectivity may be expressed statistically in terms of the degree of correspondence between scores or marks assigned to different papers by the same individual at different times. Problem 15 on Page 201 of Chapter VIII and Problem 42 in the *Work-Book in Educational Measurements*[3] afford concrete illustrations of this method of determining objectivity.

65. **Administrability.** The characteristics of a test which make for ease and accuracy in its administration are probably best expressed by the word *administrability.* While this is not one of the major criteria, it is nevertheless one that is worthy of some practical consideration. There is a very definite tendency among the makers of modern standardized tests to recognize the teachers' and the supervisors' problems of test use and interpretation by making them easy to administer and simple to interpret.

Ease of administration must be evaluated from two distinct points of view. In the first place the administration of the test from the standpoint of the pupil must be con-

[3] Greene, H. A., *Work-Book in Educational Measurements.*

sidered. Clear, simple, and direct instructions to the pupil should be printed on the test itself, and in most cases these directions should be supplemented by other directions given orally by the examiner. If some particular type of response is called for this should be illustrated clearly on the test by one or more samples.[4]

The second phase of the administrability of a test is the examiner's manual which accompanies it. This manual should contain a clear statement of the qualities measured by the test. Specific directions for administering the tests should be given so that they may be followed verbatim by the classroom teacher. It should provide the critical user of the test with an adequate explanation of the procedures by which the validity and reliability data were obtained, as well as a concise statement of the meaning of these data in relation to the tests themselves. A convenient form of answer key or stencil should be accompanied by supplementary directions in the manual concerning the methods of scoring the tests. Simple illustrations of the methods of interpreting the results should be given in the manual, and if the field is one in which a follow-up program of corrective instruction is possible, brief suggestions for such work should be given.

The better standardized tests of today provide the user with carefully formulated statements on the following types of items, all of which tend to protect the pupils and the teacher against the faulty administration of the tests, and against wrong and uncritical interpretations of the results.

1. The number of sub-parts in the test.
2. Directions for each part of the test.
3. Sample exercises to acquaint the pupil with method of response.
4. Directions for procedure at bottom of each page and at end of each test part.
5. Definite statements of time limits.
6. Directions for scoring the tests.
7. Explanations of the norms or standards.
8. Definite statements of total possible scores on each test-part.
9. Explanation and illustration of method of interpretation of results in terms of instructional needs.
10. Definite plans for remedial attack on weaknesses revealed by test.

[4] Illustrations of excellent types of directions to students and satisfactory samples for certain types of objective test exercises are given on pages 81 to 84 of this book.

The mechanical features of a test frequently operate to affect definitely its ease of use in the classroom. Such features are largely the result of the editing and printing of the test. The illustrations should be clear-cut and easily identified with the content they are supposed to amplify. Paper of good quality, preferably white bond, should be used. The page size, the length of line, and the size of type used are also mechanical features which may limit the usefulness of a test. Answer keys conveniently arranged to make for easy and accurate scoring of the test papers also contribute to the ease with which a test may be used. Class record and tabulation sheets, which by means of the method of tabulation used enable the examiner to identify and locate the pupil who makes a given low or high score, also add greatly to the utility of the test.

The administrative usefulness of norms for tests is illustrated in Exercises 19 and 20 on page 244 in Chapter IX and Problems 30 to 38 in the *Work-Book in Educational Measurements* which give much concrete experience with this phase of test interpretation.

66. Economy. Economy is certainly not one of the major criteria to consider in selecting a test, for a cheap test conceivably may be quite useless. However, many users of educational tests are beginning to be forced to a consideration of this factor by the rapid increase in the numbers of tests available for use in different fields, and the corresponding increase in the per pupil cost of testing. Real economy will not result from the indiscriminate use of cheap tests. Cost of testing should probably in the long run be computed in terms of the reliability or validity of the test per unit cost. There is a very real possibility of effecting economy in testing without reducing the quality of the content of the test by the use of specially designed test booklets which permit the tests to be used repeatedly. The development of this type of test booklet is in the future, but in it lies the most hopeful possibility of reducing per pupil cost in testing without sacrificing anything in the quality of the test content.

67. Adequacy of Test in Measuring Specific Functions. In selecting a standard test for classroom use the six criteria

previously noted for a good test should be kept in mind. Furthermore, a test should be selected because of its power to perform some specific function. If a general survey or superficial general diagnosis constitutes the purpose of the testing program the tests may be chosen accordingly. If a more detailed individual diagnosis is desired then tests adapted to provide a sufficiently detailed analysis for that purpose must be selected.

Achievement in arithmetic is the result of many skills. A general diagnosis involves simply locating pupils who have and have not achieved the general objectives set for them. The *Thorndike-McCall Silent Reading Scales* may be considered to be diagnostic in the sense that they afford a general appraisal of silent reading achievement. They, of course, do not reveal the pupil's particular limitation. They simply indicate the extent to which he is able to respond to certain types of silent reading exercises. They furnish a picture of a single type of silent reading comprehension. This vague and superficial appraisal of complicated skills is afforded by many other general survey tests in other subject-matter fields.

To ascertain the particular types of silent reading situations to which the pupil is unable to respond satisfactorily it is necessary to select a more elaborate type of analytical test such as the *Sangren-Woody Silent Reading Test* or the *Iowa Silent Reading Tests.*[5] These tests afford genuine analysis. Any less pretentious analysis of skills promises more than it can produce, and is practically useless for purposes of individual pupil analysis. Any general achievement test cannot reveal this detailed information. In most cases a study of a pupil's scores does not reveal why he failed to achieve or what particular errors he has made. In planning effective remedial instruction it is frequently helpful to secure this information by analyzing the pupil's performance. For an illustration of this type of analysis see page 252. This is going beyond measurement in the sense implied by the ordinary use of educational tests. Such analytical diagnosis is particularly profitable in the case of pupils whose achievement is markedly below standard.

[5] See list of tests and publishers on pages 585 to 595 in *Appendix.*

V. The Critical Selection of Tests

68. Test Rating Scales. In the foregoing discussion of the criteria for educational tests no attempt has been made to evaluate in a definite sense any of the items which appear to affect test quality. Numerous rating devices which weigh the various items roughly in order of importance are available. The assignment of point values to the different features of the tests is, of course, a purely subjective procedure, as is the actual use of such score cards in the hands of different individuals. It is obvious that two different individuals using such score cards could not be expected to agree closely in the point scores assigned to a particular test. However, in spite of these limitations such rating scales are of very real value to the inexperienced teacher or student because of the definite way in which attention is called to the quality features in a test. Because of the very satisfactory and detailed analysis they furnish of the factors which should be considered in a criterion for tests, the *Cole-von Borgersrode Scale for Rating Standardized Tests*[6] and the *Otis Scale for Rating Tests*[7] are both reproduced here. The student will find that time spent in the study and application of these scales in the actual evaluation of educational tests will be a most effective method of giving definite meaning to the specific test criteria discussed in earlier sections of this chapter.

Cole-von Borgersrode Scale for Rating Standardized Tests

I. Preliminary Information
1. Exact name of test
2. Name and position of author
3. Name of publisher and nearest address
4. Cost
5. Date of copyright
6. Purpose of test

[6] Cole, Robert D., and von Borgersrode, F. R., "A Scale for Rating Standardized Tests," *School of Education Record of the University of North Dakota*, Vol. XIV, No. 1. Grand Forks : University of North Dakota, 1928, pages 11-15.

[7] Otis, A. S., "Scale for Rating Tests," *Test Service Bulletin No. 13*. Yonkers, New York : World Book Company, p. 6.

II. Validity (25)
  A. Curricular (15)
    1. Exact field or range of education functions which test measures?
    2. Ages and grades for which intended?
    3. Criteria with which material was correlated?
    4. Do questions parallel good teaching procedures?
    5. How wide is sampling of important topics?
    6. What is the social utility of questions?
    7. Is test claimed to be diagnostic? (If so, proof and see VI, 5, c, below.)
  B. Statistical (10)
    1. Correlated against what outside criteria?
    2. Size of coefficient of correlation?
    3. Size and representativeness of sampling?
    4. Proof of validity of items (such as statements as to experimental tryout of items individually to determine that no large percentage is failed or passed by all pupils and that the items show a consistent increase of percentages of successes with successive age or grade levels).

III. Reliability (25)
  A. Most important items
    1. Correlated with what?
    2. Size and representativeness of sampling?
    3. Reliability coefficient
    4. The means of the distributions
    5. The standard deviations of the distributions
    6. If some other measure than the above three is given to prove reliability, what is it?
    7. Inter-correlations
  B. Less important but desirable
    1. Order of giving various forms of test
    2. Is test reliable enough statistically for individual measurement, or can it be used only for groups?
    3. •Evenness of scaling (see II, B, 4)
    4. Are pupils accustomed to this type of test?

IV. Ease of Administration (15)
  1. Manual of Directions (3)
    a. How complete and simple is the manual?
    b. Does manual control test conditions well?
    c. Typographic make-up
  2. Simplicity of Administration (8)
    a. Amount of explanation needed for pupils by examiner?
    b. Are directions to pupils clear, detailed, comprehensive?

c. Is arrangement of test convenient for pupils?
d. Are samples and "fore-exercises" given when needed?
3. Alternate forms (3)
a. Number
b. Evidence of reliability
c. Evidence of equivalence
4. Time needed for giving

V. Ease of Scoring (10)
1. Degree of objectivity — purely objective or some judgment on part of examiner?
2. Are adequate directions given — clear, equal to all emergencies?
3. Is scoring key adjusted to size of test?
4. Time needed to score one test
5. Simplicity of procedure
a. Number of processes needed to get final score?

VI. Ease of Interpretation (20)
1. Norms (6)
a. Kind — age, grade, percentile, etc.
b. Derivation — size and representativeness of sampling
c. Tentative, arbitrary, or experimental?
d. For separate parts?
e. How expressed?
2. Is class record provided?
3. Are there provisions for graphing results?
4. Is interpretation of raw scores easy or hard?
5. Application of results (10)
a. Are directions or suggestions given for application of results to benefit teaching or administration?
b. Are tests survey or diagnostic?
c. If diagnostic —
(1) Proof of diagnostic value?
(2) What principle or principles underlie construction?
(3) How many different skills, abilities, or aspects of the subject are analyzed or measured?
(4) Does the analysis of total subjects into unit abilities follow teaching practices or needs?
(5) Is the diagnosis individual or class-proof?
(6) Does the test demand tabulations of individual pupils' errors to secure diagnosis?
(7) Is a remedial program provided or suggested?

VII. Miscellaneous (5)
1. Typography and make-up
a. Arrangement of printed matter
b. Legibility of type

c. Quality of paper
d. Are test blanks free from distractions, norms, directions to examiner, etc. ?

2. Is the time required for giving as small as is consistent with reliable measurement ?
3. Is the cost in keeping with the amount, scope, and reliability of the results yielded ?
4. Is good test service provided by the publisher ?
5. Kind of new-type questions used ?

Explanation of the Otis Scale

1. Manual. (Total 7 points)
   Is the test accompanied by a properly organized and carefully prepared manual of directions explaining in detail how to give, score, interpret, and apply the results ?
2. Validity. (Total 20 points)
   Is a statement of method of validation given ? Is the method a sound one according to the evidence submitted ? Do the items tested appear to coincide with a common sense analysis of the subject ?
3. Reliability. (Total 10 points)
   Does the manual contain data indicating the reliability of the test such as reliability coefficients, or probable error of a score ? How reliable is the test ? Upon how wide a range of ability is this reliability based,— that is, on cases from a single grade or from several grades grouped together ? (The latter practice results in a spuriously high reliability index.)
4. Reputation. (Total 3 points)
   Has the test been widely used ? This is particularly important when considering the norms.
5. Ease of Administration. (Total 20 points)
   a. Little special preparation required : Little special preparation and equipment is desirable. Must the examiner have special preparation in order to give the test without error ? (4 points)
   b. Adequate detailed directions : Are all directions to pupils stated verbatim in the Manual ? (6 points)
   c. Time limits clearly stated : Are there few or many to watch ? Are the times clearly stated in manual and on tests ? (5 points)
   d. Alternate forms available if necessary : More than one equal and alternate form of tests is usually desirable. (5 points)
6. Ease of Scoring. (Total 15 points)
   a. Objectivity : Are exercises so formulated that exact answers are possible ? No judgment on the part of the scorer should be required as to correctness of response. (8 points)

OTIS SCORE CARD FOR RATING STANDARDIZED TESTS

Item	Standard Number Points	Name of Tests			
1. Manual	7				
2. Validity	20				
3. Reliability	10				
4. Reputation	3				
5. Ease of Administration (20)					
a. Little special preparation	4				
b. Adequate detailed directions	6				
c. Time limits clearly stated	6				
d. Alternate forms available	4				
6. Ease of Scoring (15)					
a. Objectivity	8				
b. Convenient form of Key	4				
c. Time required	3				
7. Ease of Interpretation (20)					
a. Types of norms	10				
b. Directions for	3				
c. Class Record Sheet	2				
d. Remedial Program	5				
8. Typography and Makeup	5				
Total	100				

b. Convenient Form of Answer Key: Is answer key provided? Is it convenient to use? (4 points)

c. Time required: Are tests easy and rapid to score or slow and tedious? (3 points)

7. Ease of Interpretation. (Total 20 points)

a. Types of norms: Are norms adapted to the use to be made of them? In elementary school tests are both age and grade norms given? In high school tests are both grade and percentile norms given? (10 points)

b. Directions for use of Norms: Does the manual provide complete directions (with illustrations) for the interpretation of the meaning of the test scores? (3 points)

c. Class Record Sheet: Is a class record sheet provided for convenience in handling and filing scores? (2 points)

d. Remedial Program suggested: Do the tests provide adequate detailed information for the development of a remedial program? Are any remedial suggestions provided? (5 points)

8. Typography and Makeup. (Total 5 points)

Is the test printed in legible as well as pleasing type on good paper so that it is easily read? Does it have a practical as well as pleasing appearance?

VI. General Rules for Administration

The following rules are common to most tests now in use. They are not intended to take the place of the directions accompanying the various tests that may be used. The directions for giving and for scoring given in the Examiner's Manual which accompanies the better tests should be rigorously adhered to, so that the tests may be given under standard conditions. It is hoped that the following suggestions may be an aid in the administration of educational tests.

69. **Preparation for Administering the Test.** Any teacher or principal who is reasonably skillful in discipline, and who will carefully follow the directions accompanying the tests should be able to administer educational tests. Unless the test directions are extremely familiar the examiner should study the manual carefully before attempting to give the test. If possible he should administer the test to some other person in order to gain further familiarity with the procedure. If this is not possible the directions should be read aloud several times so that they may be followed easily during the test. Familiarity with the directions is essential if

the standard conditions for the test are to be maintained, and comparison with the norms and standards thus be made possible.

Pupils may be tested in ordinary classroom groups or in larger groups. In case several grades are to be given the same test, time may be saved by moving the combined group into a larger room, care being taken that the seats and the desks are adjusted to the size of the pupil. It is probable, however, that the pupil will get a better score on the test if he is in familiar surroundings while taking it.

Before the tests are given out to the pupils the desks should be cleared and each pupil should be provided with one or two sharpened pencils. A number of extra pencils should be available for emergencies during the examination. The room should be quiet throughout the test. Require strict obedience and attention. No questions should be allowed during the test. Cultivate a manner which is agreeable but which at the same time suggests authority. Pupils should be made to feel "at home" in taking the test. Pupils will look forward to taking the tests without fear or nervousness if the tests are properly given and if no misconceptions with regard to the meaning and use of the results are allowed to arise.

Throughout the examination commands should be given in a forceful manner. The directions should be given slowly and with careful attention to emphasis. The voice should be just loud enough to carry to all parts of the room. The directions of the test should be followed verbatim. As far as possible prevent disturbances within or without the room which might interfere with the administration of the test. To avoid interruptions it is suggested that the teacher prepare a card carrying these words: *Testing Going On. Please Do Not Disturb.* If this card is hung on the outside of the classroom door interruptions will be less frequent.

70. **Timing.** The time limits as set in the directions for giving the tests should be strictly adhered to. This is especially true in short tests where a few minutes' or even a few seconds' error may invalidate the results. Tests should be timed to the second, or the results may not be comparable to what others get when the exact time is taken for the test.

In timing the test a stop watch is almost essential. If an ordinary watch is used, it is necessary to exercise great care to secure accurate timing. In such cases it is well to have an assistant whose sole duty is to act as timekeeper for the test. If this is impossible the following method will serve quite well:

	Hr.	*Min.*	*Sec.*
(*a*) Record time starting signal is given.....	11	18	20
(*b*) Add to this the time required for the test.		5	00
	—	—	—
(*c*) The sum is the time to signal a stop...	11	23	20

71. **Scoring.** The scoring of most standard tests is made almost wholly objective by the use of scoring keys. The answer keys and directions for scoring each specific test should be followed rigorously. Score exactly as instructed even though it may seem contrary to your best judgment. This is very important in order that the scores may be compared directly with the norms or standards which have been derived for the tests. It is best that all necessary calculations be performed twice, and that all transcribed records be checked against the pupils' test papers to see that no errors have been made. Errors of this character may readily invalidate the entire testing program.

The balance of the testing program which follows the giving and scoring of the test is largely statistical in character and on this account is discussed in Chapter VIII of this volume, *Summarizing the Results of Testing*, and in Chapter IX, *Interpreting the Results of Testing*.[8]

Discussion Exercises

1. What should be the classroom teacher's responsibility with regard to the use and interpretation of standard tests in the classroom?

[8] Practical problems in the application of the criteria for the selection of tests are given in the *Work-Book in Educational Measurements*, which is designed as a companion volume for this book. The teacher or student who has had the necessary statistical preparation may turn at once to these problems (Problems 29 to 32) in the *Work-Book*. Teachers and students who have had no experience with this type of work should wait until they have mastered the earlier problems in the *Work-Book* before attempting these particular problems.

2. Under what conditions do you believe the teacher should be responsible for the giving and administration of standard tests?
3. In your opinion, what types of tests (general mental, general achievement tests for survey purposes, or specific achievement tests for diagnostic purposes) should the teacher be encouraged to use most freely? Why?
4. What factors determine the time in the school term or year when the tests should be given?
5. Why is it thought to be desirable to have a clear-cut problem in mind in initiating a testing program?
6. What are the major criteria to use in the selection of a test for classroom use?
7. List and illustrate the major procedures for use in validating test material.
8. What are the factors which are the basis of reliability in a test? How can reliability of a test be improved?
9. Show how the principle of "sampling" affects the consistency of performance (reliability) of a test.
10. What are some possibilities for making test exercises objective?
11. Why is it important that the time limits set for standard tests be rigidly followed?
12. How do standard tests secure an objectivity and accuracy of scoring which other types of tests do not ordinarily attain?

Supplementary Project

Secure from your instructor at least three complete sample sets of standardized tests in some specified subject field. Examine the test booklets, the manuals, keys, and other supplementary material which accompany each test. Study the *Otis Score Card for Rating Standardized Tests* and the explanation which accompanies it on pages 120 and 122 of this book, and then score each of the tests on the form on page 121.

Write at the top of the columns the names of the tests you are to compare. Consider each item in the light of the explanations accompanying the score card and assign a rating to each item in proportion to its merit. For example, if no manual accompanies the test, record zero for that item. If the manual which is provided appears to be about half adequate, record a score of 3 or 4 for the item. Add the ratings for each test. The total score for a perfect test is 100 points.

Prepare a copy of this rating scale showing your rating of the tests to submit as your report on this project.

Selected References

Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching.* Boston : Houghton Mifflin Company, 1931.

Greene, H. A., Kelley, V. H., *Examiner's Manual for Iowa Silent Reading Test, Elementary.* Yonkers, New York : World Book Company, 1930.

Greene, H. A., Knight, F. B., Ruch, G. M., Studebaker, J. W., *Examiner's Manual for Compass Survey Tests in Arithmetic.* Chicago : Scott, Foresman and Company.

Kelley, T. L., *Interpretation of Educational Measurements.* Yonkers, New York : World Book Company, 1927.

Lang, A. R., *Modern Methods in Written Examinations.* Boston : Houghton Mifflin Company, 1930.

Madsen, I. N., *Educational Measurement in the Elementary Grades.* Yonkers, New York : World Book Company, 1930.

McCall, W. A., *How to Measure in Education.* New York : The Macmillan Company, 1922, pages 119-133.

Mead, A. R., "Suggestions for the Training of Teachers in the Use of Educational Measurements," *Educational Administration and Supervision,* Vol. 12 : 23-43, January, 1926.

Odell, C. W., *Educational Measurement in High School.* New York : The Century Company, 1930.

Ruch, G. M., Knight, F. B., Greene, H. A., Studebaker, J. W., *Examiner's Manual for Compass Diagnostic Tests in Arithmetic.* Chicago : Scott, Foresman and Company, 1925.

Ruch, G. M., and Stoddard, G. D., *Tests and Measurements in High School Instruction.* Yonkers, New York : World Book Company, 1927.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York : Silver, Burdett and Company, 1928.

Webb, L. W., and Shotwell, Anna M., *Standard Tests in the Elementary School.* New York : Ray Long and Richard Smith, Inc., 1932.

Wilson, G. M., and Hoke, K. J., *How to Measure* (Revised). New York : The Macmillan Company, 1928.

Woody, Clifford, and Sangren, P. V., *Administration of the Testing Program.* Yonkers, New York : World Book Company, 1933.

CHAPTER VII

THE CONSTRUCTION OF STANDARDIZED TESTS[1]

THE following problems in the construction of standardized tests are considered in this chapter:

a. Meaning of the standardization process.
b. Basic differences in standardized and non-standardized tests.
c. Securing validity in a standardized test.
d. Criteria for selection of test-exercises.
e. Preparing equal forms of the test.
f. Securing reliability of measurement.

This brief consideration of the construction and refinement of standardized tests is obviously not sufficiently detailed to afford an adequate guide to the inexperienced worker who may have ambitions to construct a standardized test. A treatment sufficiently extensive and detailed for this purpose would be a volume in itself. It should, however, be adequate to make the classroom teacher more critical of all types of measuring instruments, and at the same time a bit more appreciative of the care, the drudgery, and the expense in time and money required to produce a commercial test of a quality adequate to stand up under present day criteria.

I. THE PROCESS OF TEST STANDARDIZATION

72. Meaning of Standardization. Standardization, or the process of deriving comparative norms and standards, is commonly designated as the single important factor which distinguishes the informal from the more formal test. The

[1] This chapter is not essential to a reasonably complete understanding of the meaning and use of tests, and if desired it may be omitted by the student or instructor in introductory courses in measurements.

program of standardization demands a more critical analysis of subject matter, a more careful formulation of exercise material, a more exacting refinement of the techniques of evaluating test items, more critical standards of equality of items and of test forms, and more rigid statistical analysis. Thus real differences in the two types of tests appear. Such a point of view makes it clear that the mere derivation of a set of norms or standards for a test does not in any sense make it a standardized test. The matter of securing norms is undoubtedly the most important phase of test standardization but it is only one of several important procedures which are closely related to the standardization process.

73. **Standardized vs. Informal Objective Tests.** The real differences between the non-standardized and the standardized tests are relatively few in number. These differences, though not of particular importance numerically, may be of considerable significance qualitatively. As was shown in Chapter IV the objective test, because of the concise nature of the items comprising it, permits an extensive sampling into all of the important areas of the subject matter. This characteristic is also found in the standardized test. Both types of tests express performance in terms of a test score. Both utilize objective forms of test exercises. On the other hand, there are certain rather significant differences in the character of the two types of tests. One is designed to measure a specific bit of subject matter as organized and presented in connection with a particular school classroom and course of study. The other is essentially general in its content. The one is the work of an individual teacher, usually hurried in his preparation of the examination for a period immediately ahead, and almost certainly not widely experienced in the construction of such measuring instruments. The other is usually the product of an experienced subject-matter specialist skilled in the application of certain well-known principles of measurement or working in collaboration with a trained technician in test construction.

The standardized test may start out as a teacher-made test but before it is produced in final form it is usually necessary to subject it to a program of critical refinement. While norms and standards constitute an important feature of the

commercial standardized test, for the informal test no comparative norms or standards are necessary. In the standardized test the critical analysis of the quality of the individual items is usually very rigid. The evidence on the statistical validity and reliability of the standardized test must be available. The teacher usually determines these qualities in his informal test after it has been given to his students, if at all. These specific unlikenesses of the informal test and the standardized test may be sufficient to make it clear why it is worth while to give these two highly useful measuring instruments separate consideration in this volume.

II. Validation of Content

74. **Selection of Content.** The differing points of view regarding the selection of suitable units of subject matter for the informal objective examination and the standardized test result in real differences in test content. The classroom teacher of a school subject is, after all, the best judge of the items of information, the skills, attitudes, and abilities which have received emphasis in the course. Accordingly, the teacher of a subject should be able to select the content of an examination which will most closely parallel the material he has taught. Therefore, an informal objective test made up of content which the teacher himself identifies as closely paralleling the instructional emphasis of the course should naturally be somewhat higher in validity than one made from items selected by some other instructor, regardless of how well informed in the subject he may be.

The maker of the standardized test faces the problem of preparing an examination over content and for pupils he has not specifically taught. In order to be reasonably certain of being fair in the selection of items he must take care to include only those aspects of the subject which are almost certain to receive emphasis in any course. Naturally, this limits at the outset the selection of test content which must be general enough to fit into any school situation in which the course is taught. As a result the problems of the test-maker in the selection of valid content for his test are distinctly complicated.

The difficulties encountered in the selection of content for the standardized test depend to a certain extent upon the subject matter to be tested. If the field is one in which the skills are clean-cut and readily identified, the selection of the content may be a comparatively simple matter. In the event, however, that the subject is one in which the aims and outcomes are vague and general, the difficulties may be almost insurmountable, and there is little likelihood that any test based on such subject matter will have any generally accepted validity. The reasons for this are quite apparent. In the case of arithmetic, a subject in which the basic or fundamental facts are well known, the selection of content suitable for use in a standardized test is a relatively simple matter, and many tests of acceptable validity are available in this field. Exactly the reverse is true in the case of the social sciences. A similar condition exists in the industrial education courses. The fields in which the instructional aims are specific or highly factual lend themselves readily to the construction of standardized tests. Tests in subject-matter fields of a more indefinite nature are much more difficult to validate.

75. **Determining Validity.** In most subjects the validity of the test content is very difficult to establish by acceptable statistical or objective means. In certain fields it is practically impossible. In one subject a certain validation procedure may be effective and acceptable; in another it may be completely unsuited for use. Makers of standardized tests have resorted to many different types of validation procedures, the majority of which may be classified under the following heads:

Analysis of Curricular Content. Curricular content is represented most objectively in the textbooks and courses of study which provide the basis for the course, and in the examinations designed by the teacher for the purpose of measuring achievement in the course. The analysis of large samplings of such sources of curricular material affords for many subjects the only objective basis for determining the outcomes important enough to test. The major fallacy in such a procedure lies in the fact that it tends to perpetuate faulty and inadequate curricular content if it is present. It

does not look beyond present practices. On the other hand, the overlapping body of instructional material which is common to a large number of textbooks and courses of study, is almost certainly important content. Thus, if one were to attempt to validate and standardize a test in geography or history, one way to determine its content would be to include in it only those outcomes and skills which receive definite instructional emphasis in, let us say, five out of six courses of study or textbooks which have the confidence of a large body of school administrators.

Correlation with School Marks and Achievement. This method of validation assumes that in the long run a test has validity if the pupil's scores on it are closely related to his achievement in the subject. That is, a test in language must have considerable validity in content if pupils whose school marks in the subject are high consistently make the superior scores on the test, and those pupils whose school grades are low always make the inferior scores on the test.

Pooled Judgments of Experts. This procedure is related in many respects to the one mentioned just above. To the extent that teachers' marks are the judgments of experts the two procedures are identical. In spite of the apparent unreliability of teachers' marks for refined measurements, it is probably true that a teacher who has had a semester or a year in which to observe and form judgments on the abilities of his pupils has a superior vantage point from which to rate them on a scale of relative merit or ability. Regardless of their wishes, makers of tests are constantly forced to fall back upon this validation technique. In the long run, an educational test which with consistency picks out the pupils who, in the teacher's judgment of a specific ability, are superior and inferior, probably does have significant validity.

Correlation With Other Known Measures. This procedure may be utilized in fields in which extensive critical work in test development has already been done. There would be reason to doubt the validity of an achievement test in problem solving in arithmetic or algebra which did not show some relationship to achievement in problem solving as measured by other valid tests of these subjects. This is particularly true in the factual subjects. However, in cer-

tain general tool-skills, such as reading or language, it is not uncommon to find a very low correlation between two tests which are both obviously measures of certain types of reading abilities. The explanation for this lies, undoubtedly, in the very great complexity of such abilities. There are so many different aspects of the reading or language abilities, or so many different forms in which they may be revealed, that different types of measures of the subjects may show a reasonably independent validity and yet be only indifferently related to each other.

Returns from Widely Spaced Groups. One of the readily recognized evidences of validity in test content is the power of such material to reveal significant differences in the accomplishment of widely spaced groups. For example, a performance test for use in the ninth grade woodworking shop might be validated by administering it to groups of ninth grade pupils who have had a semester of shop work and to similar ninth grade pupils who have taken no industrial work in this field. If the test is valid in content, the differences in the scores made by the two groups should be significant. It is assumed, of course, that the pupils have actually learned something as a result of taking the semester course in shop work. This procedure is frequently used in the validation of aptitude test material. An aptitude test for the selection of prospective students in art might be validated by trying it out with groups of artists of recognized ability, and also with groups of individuals who are generally recognized as having indifferent artistic abilities. If the content of the proposed test is valid, there should be significant differences in the scores made by the two groups. Such procedures have been utilized in the validation of certain tests in which rather highly specialized skills are involved.

Rise in Percentage of Success. Test items and frequently whole tests are validated in terms of the increases in evidence of achievement which additional training brings about. It is expected that a test of general mental ability shall reveal increased scores in going from the ten-year-old group into the eleven-year-old group. A valid reading test is expected to show significant increases in scores indicative of

increased achievement as the tests are advanced from one school grade into the next. If second semester algebra students do not show a higher accomplishment on a test in this field than they produced during their first semester of work in the subject, there is reason to doubt the validity of the content of the test to reveal such differences. It is interesting to note that frequently there are test items which reveal little or no rise in success from grade to grade. This probably means that the item involves material which is very difficult to master, or that inadequate instructional emphasis is being given to it. It is possible, of course, that such a situation might arise through an item being so well learned at a lower level that no more improvement is possible. A few cases have been noted in which there was a marked fall in the percentage of success as the items were moved forward from grade to grade. The explanation here probably lies in the teaching of the wrong facts. That is, the longer the pupils studied the item the more mis-information they received. In the content subjects, items dealing with controversial issues might easily perform in this manner.

Social Utility. The validation of content in terms of social utility assumes that the course of study itself is based upon that point of view. This procedure is distinctly in line with modern theory in curriculum construction. There may be occasions when the standardized test anticipates the course of study in the adoption of the social utility point of view in subject matter selection but this is unquestionably rare.

A good illustration of test validation in terms of this point of view is found in the derivation of the *Newkirk-Stoddard Home Mechanics Tests.* These tests are based upon a course of study in Home Mechanics derived from the analysis of the activities of numerous adults in and around the home. Observations were made and records were kept showing the actual frequencies with which certain mechanical skills were called for in the maintenance of household equipment. The frequencies with which the head of the house is called upon to fix a leaking water tap, to hang a window weight, to wire a buzzer, etc., are indicative of the social importance of these activities. A course of study based on such an approach is distinctly practical. The more fre-

quently used a specific ability is the more important it is that it be taught and mastered as a part of a course of Home Mechanics. The high frequency job-activities represent the basic content for a test for such a course.[2]

Psychological and Logical Analysis. There are certain subjects in which it appears to be impossible to secure an objective or statistical basis of validation. In general, these subjects are the complex fields made up of many interrelated abilities as in the fields of language and reading. The most effective approach to the validation of the content of tests in these fields appears to be through introspective analysis. That is, a sort of arm-chair psychological dissection of the total process is made, in which as many as possible of the basic abilities are identified. In the use of this method the analysis is the initial step. Statistical refinements later make possible the selection of the qualities which are objectively measurable.

III. Criteria for Construction of Test Exercises

76. Form of Exercise as Conditioning Objectivity. The standardized achievement test would in all probability be an impossibility were it not for the development of the objective forms of test items. Objectivity in the test exercise is such an important element in the reliability of measurement which the test affords that it would be difficult to conceive of a standardized test made up of items not characterized by their objectivity. The reasons underlying this situation constitute the major content of Chapter IV of this volume.

The builder of a standardized educational test in any subject faces the very difficult problem of determining the precise form of objective technique which best fits the subject matter he wishes to test. In most cases this is a problem which can be answered only by experimentation.

77. Criteria for Selection of Exercises. The discussion of the problems of constructing informal objective tests presented in Chapter IV states certain principles which are to be

[2] Newkirk, Louis V., *Course of Study in Practical Mechanics*, University of Iowa Extension Bulletin, No. 146, Iowa City, April 15, 1926.

observed in the selection of the subject-matter content for any type of objective test. The test, in order to be useful must be valid ; it must measure what it purports to measure. For the teacher-made objective test, the local course of study is the basis for the determination of the content of the test. The standardized test, designed for use in all types of classrooms and in schools which frequently have widely varying courses of study, must be more basic in content. Thus the test-maker must arrive at some basis for selecting the generalized content of his test before he can approach the problem of developing test items covering the content.

After the instructional areas to be covered by the standardized test have been determined, the test-maker must proceed to break up the subject matter into elements representing the basic concepts. These important elements may then be stated in some objective form, the form selected depending to a certain degree upon the subject matter itself, and somewhat on the maturity level of the pupils with whom it is to be used. Frequently, it is necessary to prepare certain of these basic concepts in two or three different objective forms, selecting for final use the types which perform best under experimental conditions.

Different subjects probably demand different criteria for the selection of suitable test items. Of course, the general criterion of objectivity is assumed. The following supplementary standards are frequently applied in the selection of test exercises in standardized tests :

Uniformity of Response. Test items which appear to meet the other requirements of objectivity frequently are so stated that they allow considerable variation in response. This weakness in the test items is more likely to be found in recall items than in any other form. Accordingly, in the development of recall or completion-type items special care must be taken to so formulate the statement that it will call for a single correct response. It is usually found, however, that in spite of all possible care in this matter, the students will suggest many types of responses which are almost as exact and acceptable as any of those specified in the answer key for the test. When this happens the keys must be changed or else the items must be further revised. Other

things being equal, items which set up conditions encouraging a multiplicity of answers should be eliminated or corrected.

Freedom from Ambiguity. The elimination of ambiguity from test exercises is one of the most difficult problems the test-maker has to face. To keep ambiguity out of an exercise frequently means that he must simplify the statement. When the concept itself is simple, it is almost impossible to escape making the statement so obvious that its validity as a test item is reduced. Another aspect of the problem of ambiguity in test exercises is the fact that there are certain items which to the ignorant or poorly informed student are perfectly straight-forward and clear, but which to the critical and well-informed student involve implications which cloud the issue. That is, the better the student is informed in the field represented by the item the more likely he is to be confused by it. This is a phase of ambiguity in the statement of items which it is apparently impossible to separate from difficulty as such.

Difficulty. The difficulty of a test item is expressed in terms of the number or percentage of pupils of a certain classification who fail to respond to it successfully. In the actual scaling of test items this difficulty may be stated in a unit of measurement such as the standard deviation or probable error unit. (See discussion of these measures in Chapter VIII.) The net result of the use of such a unit is the separation of the exercises of varying difficulties from each other by such amounts as would probably result if the population on which the evaluation is based were infinite in number.

The determination of the optimum difficulty of the test items to be used in a standardized test is a problem on which there is not adequate evidence at present. Some experimental work in the measurement of spelling indicates that harder words are more effective measures when used with superior pupils. Words of lower degrees of difficulty are best used with poorer pupils. Words of approximately average difficulty are suitable for average classes. Possibly something similar to these conditions exists in other fields, but little direct evidence on the point is available.

The common practice in test construction is to attempt

to prepare exercises covering a wide range of difficulty, from very easy to very difficult. Items which are so easy that no pupil of the type on which the tests are to be used fails to respond correctly are certainly not suitable for inclusion in the test. The presence of such items would merely serve to lengthen the test without adding to the reliability of its measurement. In a similar way, items which are so difficult that no pupil is able to respond correctly should not be included in the test. Thus, items which lie at the extremes of difficulty, 100% failure and 100% success, are useless items, since no one is able to tell how far beyond these limits the difficulties may lie. An item which does not in a very direct way serve to differentiate between levels of achievement has no place in an educational test since it adds only useless dead weight.

Modern practice in the development of items for standardized tests tends to follow the policy of presenting items covering a wide range of difficulty arranged in ascending order from the very easy to the most difficult. This plan makes it possible for the lower grade or less able pupils to respond to certain items within their level of mastery without being unduly discouraged by being confronted at the outset with exercises of prohibitive difficulty. On the other hand, this plan serves to force the more able students to waste a certain amount of time working through a large number of items which are time-consuming but are not hard enough to bring out their real abilities. The allowance of liberal working periods for such tests tends to take care of this difficulty somewhat. In this way each student is allowed to work long enough to reach the spot where his abilities are taxed to the utmost. If the test items are carefully scaled in such a way that there is a gradual and continuous rise in the difficulty of the exercises a relatively small amount of time is lost by the superior pupils in working on items which are too far below their abilities. In a similar way the levels of ability of the less accomplished students are revealed quite promptly and accurately. The problems involved in the scaling or statistical evaluation of test items are quite technical and require a much more extensive treatment than can be presented in this volume.

Clues and Suggestions. One of the common criticisms of the objective test item is that it contains many suggestive elements or clues which the students soon learn to recognize as indicating a certain type of response. Unquestionably this is a significant criticism in many objective forms. In formulating objective items special care should be taken to see that undue suggestion of the truth or falsity of an item is not inherent in the form of the statement, although this is frequently very difficult to avoid. Some experience in the making of tests of the alternate response types indicates that the false or negative statements are more difficult to formulate and are more likely to contain recognizable clues. The sytematic use of such terms as *always*, *never*, *no*, *not*, as well as such prefixes as *un-* and *in-* in the negative forms of items may easily lead the student to spot them at once as clues.

Discriminative Power. Of the five supplementary criteria for the selection of test items this last characteristic, discriminative power, is probably one of the more important items. The basic function of all measurement is to place individuals along a scale of ability or achievement in approximate accordance with their real differences in ability or achievement. Such a function implies discriminative power on the part of the test. Since tests are made up of individual items it is clear that each item comprising a test must have this quality in a maximum degree if the total test is to possess it.

Discriminative power in a test or a test item means that a different quality or magnitude of response may be expected from individuals or groups possessing the abilities in question in a varying degree. Pupils with limited ability should fail the item more often than should superior pupils. This suggests a method by which the power of a test item to discriminate or distinguish between groups of pupils may be determined. The technical approach to the problem involves the computation of the statistical measure known as the *coefficient of contingency.* The practical implications of this procedure may be illustrated quite simply.

Let us assume that an experimental test has been given to a class of one hundred students having the normal range of ability in the subject. The tests will be corrected by the

use of the answer key, and the scores of each pupil in terms of the number of exercises answered correctly will be computed. On the basis of these scores the class of one hundred pupils can be divided into two groups, a superior group composed of those making scores of more than the median of the group, and an inferior group composed of those making scores of less than the median of the entire group. The next step involves an item count for all of the exercises used in the test. The number and percentage of pupils in the superior group failing on Item 1 is determined and compared

Table 5

DISCRIMINATIVE POWER OF TEST ITEMS IN PER CENT OF FAILURE BY GOOD AND POOR GROUPS

Item	Superior Group	Inferior Group
1	2.0	5.0
12	4.6	4.4
23	7.2	12.8
44	10.4	8.4
55	12.8	12.8
76	24.8	32.0
97	41.6	72.0
108	60.4	56.6
129	79.2	86.2
140	82.4	70.4

with similar data from the inferior group. A check of this type is made for all items in the test. A summary of a brief sampling of items from a typical test is given in Table 5. This table indicates that Item 1 was failed by two students per hundred of the superior group of students, by five students per hundred from the inferior group. This item thus shows a limited power to distinguish between good and poor achievement. The fact that the item is missed by such a small proportion of all pupils indicates that its difficulty is slight. Item 44, however, is missed by a smaller percentage of poor pupils than it is by superior pupils. In an item of

this degree of difficulty such a failure to register the real differences in the ability of these two groups is probably serious enough to warrant the elimination or at least the revision of the item. Item 97 appears to be an excellent item, while Items 108 and 140 should certainly be subjected to further critical study.

While this method of determining the discriminative power of test items is not precisely the one used in the critical analysis of test items for standardization purposes, it may serve to illustrate the general principles involved. Incidentally, the classroom teacher who is interested in the experimental development and analysis of informal objective examinations will find in the method illustrated a very satisfactory procedure for the anaylsis of such test items.

IV. Construction of Equal Forms

78. Equating the Forms of the Test. Two forms of an educational test are considered to be equal or equated when practically identical scores on each are made by the same individuals or by individuals of the same ability. This means that the forms of the test must be made up of test items which parallel each other closely in difficulty. In practice such close equality of item difficulty in alternate forms is obtained in one of two ways. One procedure involves the preparation of large numbers of items covering the total range of the subject matter to be tested, on the chance that in this large number of items there will be a sufficient number of practically identical difficulty to pair off in the alternate forms of the tests. This practice is frequently followed when the range of content covered by the test is much more extensive than can possibly be incorporated in the test itself when arranged in final form. This means that the alternate forms of the test may be made roughly equal as to difficulty but that there will be only an incidental paralleling of content. The writers believe that in the long run equality of test forms should imply parallel content as well as equality of difficulty of the items as represented by pupil-success in responding to the items.

When the range of content is more limited and the items

from which the selection must be made are fewer in number, it is not uncommon to prepare parallel items on certain selected and important concepts. In the one case the exercise may test the identification of the concept; in another it may test the identification of an additional phase of the concept itself, or some phase of the identification of the procedure. An illustration of this type of procedure is given in the accompanying exercises taken from the *Iowa Elementary Language Test.*[3] The items are taken from Part A of

TEST 2. LANGUAGE USAGE
Part A. Correct Word

FORM A	FORM B
1. The hat cost (1) two (2) too much	1. The hat cost (1) two (2) too dollars ....
2. We (1) set (2) sat there a long time	2. She (1) set (2) sat the dishes on the table ....
3. There was (1) no (2) know way to stop it	3. They did not (1) no (2) know how to stop it ....
4. Did you (1) here (2) hear the bell ring?	4. Why are you (1) here (2) hear? ....
5. He (1) lay (2) laid sleeping quietly	5. I (1) lay (2) laid my hat on that chair yesterday ......
6. You have two pencils. Let me have one of (1) them (2) those	6. Where are (1) them (2) those eggs you bought?
7. Do you think (1) there (2) their papers are the best?	7. (1) There (2) Their are ten boys in the game ....
8. They walked down (1) to (2) two the river	8. The (1) to (2) two boys walked together ....
9. You cannot (1) teach (2) learn a lazy boy	9. I will (1) learn (2) teach you to swim ....
10. (1) They (2) There is no use	10. (1) They (2) There are many miles from home

Test 2, Language Usage. This part of the test is designed to test the pupil's ability to decide which of two alternate words is the correct one to use in the particular sentence.

In these tests the exercises parallel each other as to content and also quite closely as to difficulty. The items are arranged in ascending order of difficulty as represented by the number of errors made by pupils in responding to them in

[3] Ballenger, H. L., and Greene, H. A., *Iowa Elementary Language Tests.* Minneapolis and Philadelphia: Educational Test Bureau, 1929.

the preliminary experimentation with the material. They are also arranged in such a manner that the two forms represent almost exactly the same difficulty as a whole, as well as almost parallel difficulty at any given point in the test.

In the arrangement of the items used in the above illustration, the two easiest items were found to be the ones assigned to first position in the two forms. However, it may have happened that the item used in Form A was missed by four students per hundred while the one used in Form B was missed by six per hundred. Item 2 was next in relative difficulty but it happened that one form of it was missed by seven pupils per hundred while the other statement was missed by nine per cent of the cases. Accordingly, the form of the item missed by the large per cent of the cases was assigned to Form A of the test to compensate for the fact that Item 1 in Form A was slightly too easy. This method of shifting and balancing the items for the two forms of the test results in a roughly scaled test of two or more forms composed of items likely to be failed by approximately the same percentages of cases. The accuracy of this method of equating test forms depends to a large degree upon the extent and the representative nature of the sampling of pupil-responses used in the preliminary evaluation of the items.

V. Critical Aspects of Reliability

79. **Reliability of Measurement.** Reliability has been described as the expression of the consistency with which a test performs. In the critical analysis of tests, reliability is commonly expressed in terms of the correspondence between results obtained from successive administrations of the instrument. In the development of the critical data on the reliability of a standardized test there are three possible approaches. In the first place, a single long form of the test made up of exercises scaled but not divided into equal forms may be given to an adequate group of pupils. The scores made by these pupils are then compiled for the odd-numbered items and the even-numbered items comprising the test. This has the effect of splitting the test into two

roughly equivalent parts and provides two sets of scores which may be correlated. The resulting coefficient of correlation indicates the degree of correspondence between the two chance-halves of the test. The reliability coefficient which would be expected for a test as long as the two halves combined is found by the application of the *Spearman-Brown Prophecy Formula,* an arbitrary formula designed for that purpose. This method of determining the reliability of educational tests has been very popular in the past, since it involves a minimum of labor and expense. A single administration of a fairly long form of the experimental test would provide enough data for the necessary calculations. Recently, however, there has been a reaction against such a procedure a part of which no doubt comes from the fact that the Spearman-Brown formula is somewhat arbitrary and may not always prophesy in exact accordance with the facts.

The second procedure followed in determining the reliability of a standardized test is to administer a single form of the test to a group of pupils a second time. This also provides the essential data for determining the correspondence of the two sets of scores. It is limited somewhat by the fact that previous experience of the pupils with the same test content tends to cause significant increases in scores which may be traced mainly to practice effect. There is, of course, some practice effect when different forms of the tests are used, but in this case the content is at least not identical and the score-increases are not so significant.

The method of finding the reliability coefficients for tests which appears to be most acceptable today calls for the administration of two approximately equal alternate forms of the test to the same pupils under controlled conditions. In securing the necessary pairs of scores in this way it is probably better to offer the two forms of the test in the same order so that any practice effect which operates may take place uniformly. The scores made by each pupil on Form A and Form B may then be correlated and the resulting coefficient is the coefficient of reliability (r_{AB}).

Some experience in the development of standardized tests in certain school subjects has indicated to the authors that in fairly long forms of tests the use of the chance-half

method tends to produce a spuriously high reliability coefficient. For example, the r_{11} of a long silent reading test stepped up by the Spearman-Brown formula[4] resulted in a reliability coefficient of .96 for the sixth grade based on approximately 600 cases. On the other hand, the rest r_{AB} (correlation of Form A with Form B) which was found was .90. It is the judgment of the authors that students interested in making a critical analysis of the reliabilities of standardized educational tests should do so on the basis of the correspondence of the scores on one form of the test with the other form. The resultant coefficient is more likely to be conservative, safe, and free from factors tending to force the relationships.

When the coefficient of reliability of a test is known it is then possible to express the reliability of measurement of the test in terms of the Probable Error of a test score. The probable error is a measure of the probable differences between the results obtained by the use of a single form of the test and the average of a very large number of repetions of such forms (or in other words, the "true score"). In the long run the higher the reliability coefficient of a test, the smaller the probable error of the score expressed in terms of the test scale. The data given in Table 6 show the reliability coefficient (r_{AB}), and the Probable Errors of Scores (P. E. Score) for the total comprehension scores (Tests 1 to 5 inclusive) on the *Iowa Silent Reading Tests; Advanced.*[5]

TABLE 6

RELIABILITY COEFFICIENTS AND P. E. OF SCORES
IOWA SILENT READING TESTS; ADVANCED

Test	Highest			Lowest		
	r_{AB}	P.E.	Grade	r_{AB}	P.E.	Grade
Total Compr.	.96	4.98	9	.95	5.73	12

[4] See *Glossary* for definitions of new terms.

[5] Reproduced with the permission of the publishers, World Book Company, Yonkers, New York, 1933.

80. Reliability of the Norms. Norms provide the user of a standardized test with the basis for a practical interpretation and application of the results. Naturally, unless the norms which accompany a test reflect a representative picture of the type of accomplishment to be expected they are useless and they render the test itself useless. The importance of test norms is pointed out in an extensive discussion of the subject in Chapter IX of this volume.

The type of norm or standard to accompany a standardized test depends to a certain extent upon the subject-matter field covered by the test and somewhat upon the range of the talent which it is designed to measure. A test intended to measure a tool skill such as silent reading ability naturally covers a wide range of pupil ability. It must be designed for use in a considerable number of grades, for example. It may be used in the elementary school grades or the secondary and college years. If for use in the elementary school, both age and grade norms will be found useful. If for use in the high school or college levels, age norms are of little use. Norms for the middle or end of each year of work are useful. Percentile norms are also particularly valuable in the interpretation of individual student scores, particularly at the levels at which age norms are lacking in reliability. Percentile norms are equally useful for inter-grade comparisons in the upper grade levels. Thus the content and general utility of the test, to a certain degree, conditions the type of norm which should accompany it.

One of the major problems in the derivation of norms for standardized tests is the matter of the reliability of the norms themselves. Possibly the statement of this problem would be made clearer if the word *universality* were substituted for reliability in the foregoing statement. Reliability implies consistency; but universality reflects the generally representative nature of the norms. An otherwise splendidly made test may be limited in its usefulness through the fact that the norms are not sufficiently representative. It may be that it is hopeless to expect to produce norms which are so generalized that they represent suitable bases of comparisons wherever they may be found or used. However, the only hope lies in one of two directions. One is to sample

so widely in the possible areas of population likely to use the test that practically every type and character of pupil and school situation is included. The other is to recognize the practical difficulties in the way of making a general norm fit all types of situations and select the population used in the derivation of the norms to represent deliberately chosen types of school situations.

It would seem to be impractical to expect that pupils from small school systems with little or no laboratory equipment should achieve at a level comparable to that demanded of students from schools with large well-equipped laboratories. The solution of this problem may lie in the establishment of representative norms for different types of courses and schools.

An illustration of a deliberate attempt to secure greater reliability and meaningfulness in the norms for a standardized test is seen in the norms recently made available for the *Iowa Silent Reading Tests; Elementary Test.* In these norms, recognition was taken of the fact that many of the elementary school systems of certain Southern states operate on a seven-year elementary school organization. This means that norms designed for the standard eight-year elementary school organization of the remaining states would be lacking in meaning and usefulness for these Southern schools. Accordingly, special samplings were taken and special tabulations were made to provide norms suited to the different types of school organization.

A further practical problem faced by the test-maker is the matter of determining how large a population must be sampled before he has a basis for a reliable norm. Statistically, it should be possible to determine this limit of required sampling by computing some measure of the reliability of the mean of the distribution. Some experience in the development of such norms indicates that a more practical method is the one which may be called the method of cumulative sampling. In this procedure additional samplings are added to the distribution until a place is reached where the means (averages) cease to vary when more cases are added. Naturally a norm based on one case would be unreliable. The addition of one more case would help but

it would still lack reliability. A point is finally reached, however, where the addition of more cases does not make any significant change in the value of the consecutive arithmetic means.

Discussion Exercises

1. What three basic differences appear to distinguish the informal objective examination from the standardized test?
2. Show how the process of standardization involves much more than the mere establishment of norms for a test.
3. What specific differences in points of view condition the selection of subject-matter content for the two types of objective tests?
4. Are there really two somewhat different levels of test validity?
5. Evaluate the eight suggested methods of validation of test content as they would be applied in some specified field of subject matter. Would all be equally effective as validation procedures? Why?
6. Enumerate the criteria which are suggested as the basis for the selection of the type of testing technique to be used in a test.
7. Show how discriminative power in a test item contributes to its validity.
8. How would you account for the fact that test items are sometimes found which are missed more frequently by the superior members of a class than they are by the inferior members? Would such an item be a good one to use in a test?
9. What reasons can you put forward for the preparation of a number of equal and interchangeable forms of a standard test?
10. How are equal forms of standardized tests secured?
11. Is reliability in an educational measuring device a quality which is wholly confined to the test itself? To what extent do you think that unreliability in measurement is contributed by human elements?
12. Upon what are statistical expressions of the reliability of standardized tests based?
13. Which would you expect to show the highest relationship, pairs of scores resulting from the repetition of the same form of a test, or pairs of scores resulting from the use of two equal and alternate forms of the test?
14. What factors appear to determine the type of norm which should be supplied with a standardized test?

15. What does the term "universality" imply with respect to test norms? Is it synonymous with reliability?

Selected References

McCall, W. A., *How to Measure in Education.* New York: The Macmillan Company, 1922.

Monroe, W. S., *Introduction to the Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923.

Odell, C. W., *Traditional Examinations and New-Type Tests.* New York: The Century Company, 1928.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929.

Thorndike, E. L., *Mental and Social Measurements.* New York: Teachers College, Columbia University, Bureau of Publications, 1913.

CHAPTER VIII

SUMMARIZING THE RESULTS OF TESTING

This chapter gives consideration to the following points in the interpretation of test results:

a. Statistical procedures needed in summarizing test results.
b. Tabulation of test scores.
c. Common measures which express typical achievement.
d. Methods of computing these measures.
e. Common measures of spread or variation.
f. Inter-relationships of test scores and abilities.
g. Methods of using test scores to place pupils in order of ability.

It is a matter of common knowledge that a wide range of accomplishment may be expected from the different individuals in a given class. This means that scores representing objective measures of achievement in the classroom will vary widely. Since the human mind is not able to grasp and hold numerous unlike facts in isolation, it is apparent that some methods of summarizing these widely varying results so that they may accurately and simply be described, are necessary. Summaries and descriptions of this type are statistical problems, but they need not disturb the student or teacher, for after all most of these elementary statistical procedures are simple. The main difficulties are the learning of a new and different type of vocabulary, and the revival of a few relatively simple arithmetical skills.

The use of statistical methods in the analysis of test results is directly in line with good scientific technique. Scientific methods involve first the collection of facts. Within the limits of accuracy of the tests used, the test scores may be said to represent facts. The second basic step in scientific procedure calls for the classification and organization of the facts. The simple statistical practices of grouping and

tabulating data are utilized for this purpose. The next step in scientific methods involves the further reduction and analysis of the data. The common statistical procedures such as determining measures of central tendency, variability, and relationship are required at this point. Clearly very few significant generalizations have been reached without the use of some method of bringing the varying and often conflicting masses of data into focus. As a matter of fact, the use of simple statistical procedures in the interpretation of test results may be likened in many respects to the use of a double-convex lens for the purpose of bringing parallel rays of light into focus.

The most important statistical techniques from the standpoint of the frequency of their use in the classroom are: (1) a knowledge of how to classify and tabulate data; (2) a knowledge of how to compute the common measures of central tendency; (3) a knowledge of how to express the variability or spread of the data; (4) a knowledge of how to express the relationship between two or more groups or arrays of data; (5) a knowledge of how to utilize simple graphic methods in the presentation of facts; and (6) a knowledge of how to secure and use standards, norms, and derived scores for purposes of comparison and interpretation of test results.

The discussion of these items constitutes the major portion of this and the following chapter. This relatively large amount of emphasis is given to these points for two reasons: (1) successful and satisfactory work with tests can be expected only when the person using them is adequately equipped to understand and interpret them. Such abilities are dependent upon a reasonable mastery of these elementary statistical techniques. (2) Current educational literature in practically all fields is literally filled with the terms and the techniques discussed here. Reports of progress in education are dependent upon statistical methods. If the teacher and the student are to keep up-to-date educationally, they must develop the ability to read with understanding and to apply practically the statistical discussions in current educational literature.

I. Classification and Tabulation of Test Scores

81. Need for a Method of Grouping Data. The very fact that people are unlike physically and mentally gives rise to the need for statistical methods in psychology and education. This is illustrated by the history test scores given in Table 7. Pupil A, an eleventh grade pupil, made a score of 52 points on a certain history test. Pupil J, in the same school grade, made a score of only 20 points. An examination of the data in this table shows that other pupils made scores between these two extremes. If this were a large class and the range of scores (the difference of the highest and the lowest scores) were greater it would be very difficult for a teacher to secure a clear picture of the actual achievement of this class as represented by these scores. As a matter of fact the arranging of scores in order of size is one of the very simplest steps in summarizing test results for interpretation. It enables the teacher to pick the highest and the lowest scores easily, and also after some practice to pick from the scores the single one that seems to be most typical of the group.

Table 7

SCORES ON A HISTORY TEST BY AN 11TH GRADE CLASS

Pupil	Score
A	52
B	50
C	47
D	41
E	36
F	36
G	30
H	27
I	21
J	20

The relatively simple practice of arranging test scores in order of size from highest to lowest or the reverse is helpful,

however, only when the number of pupils in the class is not large. When there are ten or more it becomes more helpful, perhaps even necessary, to arrange the scores into groups more compactly so that they may be easily read and comprehended. This process of grouping the cases into the form of a table is called *making a frequency distribution* and the resulting table is called a *frequency table.* The need for this method of grouping scores is apparent after examining Table 8 which gives the scores of thirty-seven 9th grade pupils on an algebra test. It is impossible to grasp the significance of such a large number of widely scattered test scores without resorting to grouping.

Table 8

ALGEBRA TEST SCORES OF THIRTY-SEVEN 9TH GRADE PUPILS

72	71	68	68	66	66	50	60	48	46
46	45	43	25	52	53	53	54	55	55
63	58	57	40	38	40	36	34	30	32
30	28	24	70	60	64	40			

The method of grouping test scores is similar to that followed by postal clerks in the distribution of letters and other mail in a large railway postal terminal. Mail designated for certain sections of the country or for certain large centers from which it is re-distributed is thrown into the proper mail pouches. Each pouch may, however, contain mail addressed to a number of postoffices in the same section of the country. The number of pouches required depends upon the number of pieces of mail to be distributed and also upon the size of the section of the country which can be most efficiently served by a given pouch. Increasing the number of pouches naturally increases the labor involved in sorting the pieces of mail, but at the same time it increases the accuracy of the distribution. Mail in Chicago might be sorted into two classes — eastbound and westbound. This would introduce a large error since not all sections of the country would be effectively served by this rough classification.

82. **Steps Involved in Preparing a Frequency Table.** The purpose of the foregoing illustration is to give the reader

a clear conception of the principles and problems involved in grouping test scores into a frequency table. In the data given in Table 8, page 152, the scores range from the low score of 24 to the high score of 72. The process of grouping data involves the following three steps:

(1) *The determination of the range of the scores.* This is similar to determining how far apart the sections of the country are which are to be served by the postal terminal station. This aids in determining how many mail pouches must be used.

(2) *The determination of the number and the size of the groups in the classification.* If a given mail pouch is to serve a great many minor postal stations a larger type of pouch and a smaller number of pouches may be used than if each minor postoffice has its own individual pouch.

(3) *The tabulation of the scores.* The pieces of mail can not be sorted accurately and effectively until the number of mail pouches and the exact sections of the country served by each have been determined. After this has been done the trained postal clerk can accurately and rapidly sort the pieces of mail into their proper pouches.

The result of applying these three specific steps to the test scores in Table 8, page 152, is given in Table 9, on page 157. This is representative of the type of frequency table around which the following discussion will center.

83. Determining the Range of Scores. The *range* of scores is found by subtracting the smallest score from the largest score. In the case of the scores given in Table 8, page 152, the range is the difference of 24, the lowest score, and 72, the highest score. The range is therefore 48. The range is useful in determining the number of class-intervals or steps to be used in making the frequency table. (See Table 9, page 157.)

84. Determining the Number and Size of Class-Intervals. Since the scale of a frequency table must include the entire range of scores present, the size of the class-interval used depends upon the range of the scores, and upon the number of such class-intervals, or steps, which it seems wise to use in the table. No special rule relative to the number of steps or class-intervals to be used in a frequency table can be stated.

However, it is probably safe to state that it is unwise to group data into fewer than twelve or fifteen class-intervals because of the greater *error of grouping* as the number of steps is decreased. This point was suggested in connection with the illustration of the sorting of mail in the terminal station. If all mail westbound from Chicago were placed in three pouches consigned to St. Paul, St. Louis, and Omaha, considerable injustice would be done to other sections of the country. The more the number of pouches is increased the more accurate the resulting routing of the mails—up to the point where so many separate pouches are used that they can not be handled economically. In a similar way there is a number and size of class-interval which seems to be best for each distribution.

It is usually unnecessary to use more than twenty to twenty-five steps or class-intervals in making the frequency table in practical work with educational test data. The use of more than twenty-five steps involves somewhat more labor in making the distribution, and experience shows that where there are enough cases to warrant making a frequency table the use of more class-intervals does not greatly affect the error of grouping involved. The main idea of grouping the scores into approximately fifteen to twenty steps as suggested is to organize the scores into sufficiently small groups that they may be thought about effectively and yet not classify them into so few groups that important differences are covered up or significant errors are introduced.

When the two factors (1) *the range of the scores*, and (2) *the approximate number of class-intervals* which is to comprise the table, are known, the size of the class-interval itself can be determined. The range of the scores found by taking the difference between the largest and the smallest scores in the array, gives a suggestion as to the number of units which may be used in each step or class-interval, assuming that the table is to have approximately 15 steps. Let us refer again to the scores given in Table 8, page 152. The range is from 72 to 24 points, or 48 points. If the range is divided by the number of class-intervals desired the resulting quotient indicates that the size of the intervals in this case should be slightly more than three units if only fifteen steps

are used. However, the number of steps is somewhat flexible, so that for these data the use of class-intervals of three units each is desirable. (See Table 9, page 157.) Experience with the tabulation of test scores indicates that class-intervals of one, three, five, etc., units are more convenient for use than steps of two, four, six, etc., units. Intervals of one, three, five, etc., units simplify later calculations by making the mid-points of the steps conveniently whole numbers. By using intervals with two, four, and six units the mid-points are not only not whole numbers, but the limits of the class-intervals themselves are also fractional. This point will be discussed further in a later section in this chapter.

85. Determining the Limits of the Class-Intervals. The next step in setting up the class-intervals for a frequency table is that of establishing the *limits* of each interval. In order to make this more easily understood a few terms must be defined. The *class-interval* or *step* is the group, or compartment within the limits of which given scores are assigned. The *mid-point* of the step or interval is a point mid-way between the upper and the lower limits of the step. For convenience in tabulation it is found desirable to choose the limits of the step in such a way that this mid-point is a whole number.[1] This, of course, makes it necessary that the upper and lower limits of the step be fractional values whenever odd-sized steps are used, but this is thought by many statisticians to be desirable also for the following reason : In most educational measurement there is the possibility of fractional scores, or of work on exercise units partially completed. It is thought that a score of 42 may as often really represent a score of 41.5 as it does a score of 42.5. The most likely score representing scores varying from 41.5 to 42.5 is of course 42. This method has the additional merit of giving a natural location on the scale

[1] At this point attention should be called to a rule of tabulation which if followed by members of large classes will greatly simplify the work of checking the papers. The rule is as follows : Select as mid-points for the class-intervals values which are divisible by the number of units per step. For example, in frequency tables using three unit steps the mid-points should be multiples of three, as 18, 24, 27, 30, etc. If every student follows this rule the tabulations will all be made in the same manner and may be checked from one answer key. It is from this standpoint alone that this suggestion is made and not because of any statistical merit which it may have. The problems in this book are worked on this basis.

to all scores expressed in whole numbers. A further advantage is that it makes it possible to compute a median (explained and discussed later in this chapter) from grouped data which is practically identical with the median found from ungrouped data.

It is now possible to proceed to set up the class-intervals for the frequency table. In the case of the data given in Table 8, page 152, it is clear that the table must extend over a range of 48 points. To do this the table must provide class-intervals or steps high enough on the scale to accommodate scores of 71 and 72. At the same time steps must go low enough on the scale to accommodate scores as low as 24. The procedure at this point is, therefore, to begin at one end or the other of the distribution and build up the steps systematically. For example, if we start at the bottom with the lowest score in Table 8 it will be necessary to select as the mid-point of the step the value 24 in order that it may go low enough on the scale and at the same time be divisible by the number of units in each step. Since the interval is comprised of three units with 24 as the mid-point, all scores of 23, 24, and 25 will go in this interval. In order, however, that all possible values of 23, 24 and 25 may be accommodated within the interval the limits of the interval must be established at 22.5–25.5. The accompanying diagram (Figure 4) shows this.

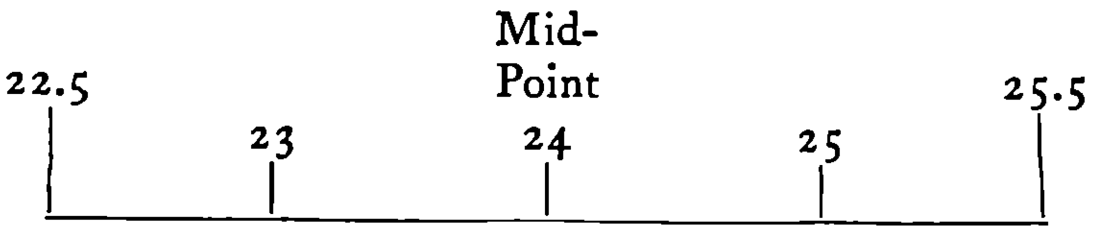


FIGURE 4. MID-POINTS AND LIMITS OF CLASS-INTERVAL

The results of applying these principles to the construction of a frequency table are shown in Table 9 on page 157 which is the frequency distribution of the Arithmetic Scores given in Table 8 on page 152.

86. Tabulating the Scores. The tabulation of scores is nothing more than a filing or pigeon-holing process which assigns to places within the same class-intervals scores which fall within the limits of these intervals. Table 9 shows

ıat there were two scores which fell within the limits of the ıterval of 70.5 to 73.5; that there were three scores ı the interval 67.5 to 70.5, etc. The appearance of a score ı an interval is usually indicated by a short vertical mark. 'he presence of five scores in a step is indicated by four ver- cal marks (or tally marks) crossed off by a fifth mark drawn

TABLE 9

SCORES GIVEN IN TABLE 8 GROUPED IN INTERVALS TO FORM A FREQUENCY DISTRIBUTION

(1) Class Intervals	(2) Mid-Points	(3) Tabulated Scores	(4) Frequency (f)	(5) Deviation (d)	(6) fd	
70.5–73.5	72	//	2	+7	+14	
67.5–70.5	69	///	3	+6	+18	
64.5–67.5	66	//	2	+5	+10	
61.5–64.5	63	//	2	+4	+ 8	
58.5–61.5	60	//	2	+3	+ 6	
55.5–58.5	57	//	2	+2	+ 4	
52.5–55.5	54	~~////~~	5	+1	+ 5	+65
49.5–52.5	51	//	2	0		
46.5–49.5	48	/	1	−1	− 1	
43.5–46.5	45	///	3	−2	− 6	
40.5–43.5	42	/	1	−3	− 3	
37.5–40.5	39	////	4	−4	−16	
34.5–37.5	36	/	1	−5	− 5	
31.5–34.5	33	//	2	−6	−12	
28.5–31.5	30	//	2	−7	−14	
25.5–28.5	27	/	1	−8	− 8	
22.5–25.5	24	//	2	−9	−18	−83
Total Number of Cases (N)...... 37			37			

Total Frequencies.................... 37
Mean, Ungrouped Data 49.78
Mean, Grouped Data 49.53
Median.......................... 51.75
Range........................... 48
Upper Quartile.................... 61.13
Lower Quartile.. 38.44

diagonally through the four, as in the case of the interval where the mid-point is 54 in Table 9. The number of check marks in each step is then usually totaled and entered under the column marked *frequency*.

87. Summary of Steps in Making a Frequency Table. The classroom teacher will very often find the construction of a frequency table unnecessary in his work with tests since he usually works with small numbers of cases and can check the scores from the papers themselves. However, there are many occasions when the frequency distribution is necessary. It is an effective way of recording and preserving the results of using tests in the classroom. It makes possible a number of short cuts in the calculation of certain statistical measures useful in interpreting test results. The methods by which these measures are computed are given in succeeding sections of this chapter. This section summarizes in concise form the steps necessary in making a frequency table.

(1) Determine the range of the scores by subtracting the smallest score in the series from the largest score.

(2) Determine from the range of the data the number of class-intervals which should be used. No special rule covering the number of class-intervals can be given. However, it is probably safe to state that it is usually unwise to group data into fewer than twelve to fifteen steps and probably unnecessary to group data into more than twenty-five steps.

(3) Use a class-interval of one, three, five, etc., units in preference to one of two, four, six, etc., units. Table 10 may be a helpful guide in your first experience with this type of work.

(4) Select as mid-points of the class-intervals values which are multiples of the size of step used. If a step of three units is used, select mid-points which are divisible by three, etc.

TABLE 10

RELATION OF RANGE AND SIZE OF CLASS-INTERVAL

For a Range of	Use a Class-Interval of
25 or less	1
26 to 69	3
70 to 125	5
126 to 175	7
176 or more	15

(5) Set up the mid-point values of the intervals required and establish their limits clearly, using the fractional values as limits.

(6) Tabulate the scores determining the frequency of scores in each class-interval and the total frequency of scores in the entire table, as illustrated in Table 9, page 157.[2]

PROBLEM-EXERCISES IN TABULATING TEST SCORES[3]

Problem 1

Physics Test Scores of 21 Twelfth Grade Pupils.

a. Arrange or rank these scores in descending order:
43, 15, 18, 43, 20, 43, 26, 42, 29, 40, 29, 38, 36, 36, 30, 35, 32, 35, 32, 32, 32.

b. What is the largest score made on this test by any pupil in the class?

c. What is the lowest score made on this test by any pupil in the class?

d. What is the range of these scores?

e. If a frequency table of ten steps is decided upon how many units will there be per step?

f. What will be the limits of the class intervals used?

g. Tabulate these scores in a Frequency Distribution. Choose your own size of step and number of class-intervals. Remember to choose steps with mid-points divisible by the size of the class-interval. Do all of your work on the left half of a sheet of paper and *preserve it for use in connection with other problem work.*

Problem 2

Reading Test Scores of 41 Ninth B Grade Pupils.

38, 36, 46, 71, 61, 49, 40, 67, 44, 52, 58, 51, 57, 21, 41, 32, 50, 80, 39, 77, 46, 57, 65, 58, 61, 72, 55, 57, 57, 61, 38, 66, 41, 66, 48, 33, 49, 36, 46, 34, 44.

a. Determine the range, set up a frequency table of a suitable number of steps, and tabulate the above scores. Do all of your work on the left half of a sheet of paper and *preserve it for further use.*

[2] Extensive practice on the tabulation of data is necessary before the student will find it easy. For this purpose the student is referred to *Work-Book in Educational Measurements* by H. A. Greene. This *Work-Book* consists of a series of forty-five practical problems in the tabulation, analysis, and interpretation of test scores. At this point see particularly Problems 1, 2, 3, and 4 of Work-Unit I.

[3] For the purpose of supplying the interested reader with an opportunity for additional experience in interpreting the various phases of test results, problem-exercises of this type are distributed throughout this chapter, and Chapter IX.

Problem 3

Language Test Scores of 31 Tenth Grade Pupils.

50, 34, 34, 34, 45, 34, 33, 41, 33, 33, 40, 32, 32, 40, 32, 32, 37, 11, 14, 37, 37, 37, 29, 10, 31, 30, 24, 25, 29, 29, 29.

a. Determine the range, set up a frequency table with a suitable number of steps, and tabulate the above scores. Do all of your work on the left half of a sheet of paper and *preserve it for further use.*

II. Measures of Central Tendency

The grouping of test scores into frequency tables is one step in the process of condensing them so that they can be analyzed and interpreted. However, a further step must be taken before it is possible to describe the table itself. Some single term or value which is representative of the entire table must be found. Since these values, which may be taken to represent an entire distribution of scores, are usually found near the center of the data when arranged in order of size they are commonly called *measures of central tendency.* The three common measures of central tendency are: (1) the arithmetic mean; (2) the median; and (3) the mode. Of these three measures of central tendency the median and the arithmetic mean are used almost exclusively in educational measurements, and are accordingly the only ones emphasized in this discussion.

88. Computing the Arithmetic Mean — Ungrouped Data. Practically everyone uses and knows how to find the so-called average or *arithmetic mean.* It is commonly defined as the measure resulting from dividing the sum of the measures in the distribution by the number of measures. In the case of the scores shown in Table 7, page 151, the arithmetic mean is found by taking the sum of the scores earned by all of the pupils involved and dividing this by the number of pupils in the class. The sum of the scores earned by pupils A, B, C, etc., is 360. There are ten pupils in the class, therefore 360 is to be divided by 10. The result, the arithmetic mean, is 36. The value of this measure lies in the fact that it lends itself to describing by means of a single term a group of widely varying scores or measures. It expresses in very com-

pact form one specific fact about the scores in which each single score has a part. On this account it is one of the basic statistical measures of central tendency.

89. Computing the Arithmetic Mean — Data Grouped in Frequency Distribution. The arithmetic mean can also be readily found from a frequency distribution. In order to make the procedure somewhat more definite it is advisable to redefine the term *arithmetic mean.* When considered from this point of view *the arithmetic mean is defined as a point on the scale, such that the sum of the deviations of the values larger exactly equals the sum of the deviations of the values smaller than it is.* Expressed in physical terms it may be considered the point at which the fulcrum must be placed in order to balance the scale when it is considered as a beam of regularly increasing thickness or density. For example, a beam of a certain length, two inches in thickness and varying in width from two to seven inches, can be balanced by placing a fulcrum at the proper point. This point may be determined experimentally or by mathematical calculation. Without regard to the method employed, the fulcrum must be so placed that the moments of the forces on one side

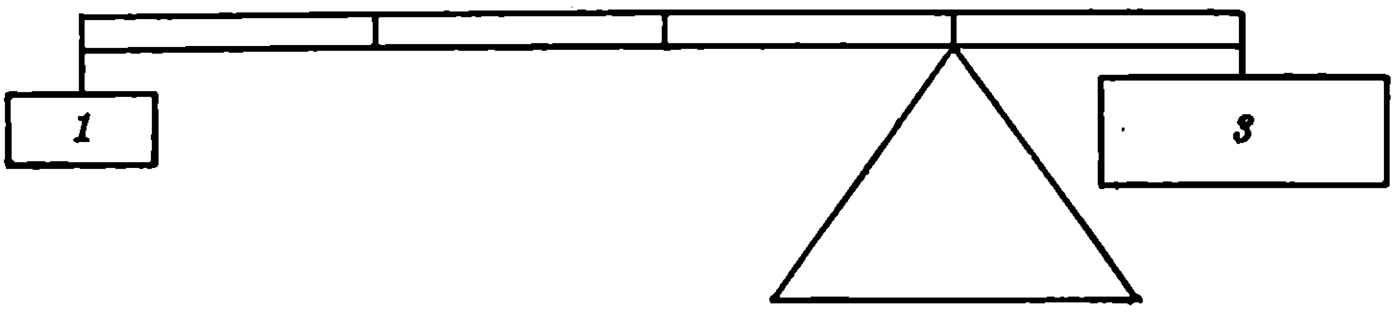


FIGURE 5. ILLUSTRATING PRINCIPLE OF MOMENTS OF FORCES

are exactly equalled by the moments of the forces on the other.[4]

The parallel between the physical lever and the mathematical calculation of the arithmetic mean is quite close. The problem in each case is to balance the forces on either side of a point to be determined. In case the physical lever is out of balance the correction is made by moving the fulcrum in the direction of the heavier end until equilibrium is estab-

[4] Those who have forgotten the meaning of the term *moments of force* will be reminded by the statement that a force of three pounds resting on a lever one foot in length is exactly equalled by a force of one pound resting on a lever three feet in length. (See Figure 5.)

lished. In calculating the arithmetic mean a sort of trial balance is taken. If the moments of force are too great on one side the point of rotation is similarly moved in the direction of the heavy end until the difference between these two forces becomes zero. This may be aptly illustrated by taking a procedure which classroom teachers have undoubtedly frequently used. It is necessary to obtain the average of five class marks for various periods through the school year in arriving at the final mark. These grades may be taken for illustration as 93, 90, 89, 88, 86. By inspection it may be seen that 89 is approximately the correct mean. The 90 is one point too large, the 93 is four points too large. In a corresponding way 88 is one point too small and 86 is three points too small. The total of the differences above the mean is five and the total of the differences below the assumed mean is four ; therefore the assumed mean of 89 is too small by the amount of this difference (1) divided by the number of cases (5). This is equal to .2, and the mean is 89.2. This checks exactly with the mean found by the method of totaling the measures and then dividing by the number of measures.

This principle may be applied almost directly to the calculation of the arithmetic mean from the frequency table. A sort of trial balance is taken by looking at the array and assuming as a mean some value on the scale which approximates the true mean. As a matter of fact it is not necessary that this assumed point should be even approximately near the true mean. Any point on the scale could be assumed with equal accuracy. For the sake of illustration refer to Table 9, page 157, where the step marked "0" in the deviation column will indicate the point taken as the assumed mean. This point is called *zero* since all the measures in that particular step are assumed by this process to deviate zero from the mid-point of the step. Deviations in terms of steps which are representative of the intervals on the scale, comprising in this case three units, are laid off above and below. The steps in the ascending order of magnitude in the scale are designated as plus. Those in the descending order are designated as minus.

The total magnitude of forces on the upper half of the

scale is found by multiplying each frequency by its corresponding step deviation and taking the sum of these products (marked in Table 9 as *fd's*). The sum of the plus *fd's* in this calculation is 65. The sum of the minus *fd's* is 83. The algebraic sum of these two quantities is —18 (65 and —83 = —18), which expresses approximately the force tending to rotate the scale in the direction of the greater magnitudes. In order to bring about an exact balance of these two forces, the assumed mean which was selected at the outset must be moved slightly in the direction of the heavier end of the scale which is in this case in a minus direction. This is illustrated graphically in Figure 6. Since there are 37 measures in the distribution and each of the 37 measures contributes to the resultant force of 18 units, the average correction in this case is the result of dividing —18 by 37 which gives a correction of approximately —.49 steps.[5] Since each step is based on three units it is necessary to multiply —.49 by 3 in order to turn it into scale units. The resulting value of —1.47 units represents the exact amount by which the assumed mean must be corrected. The sign of this value resulting from the algebraic addition mentioned above is negative, therefore, the correction must take place in the negative direction.

When the assumed mean was taken, the step represented by the values 50, 51, and 52 on the scale was tentatively accepted as that point. Since there are possibilities for a number of values to fall within this step in making the frequency table it becomes necessary for some one value to be assigned as most nearly representing the step. It is the practice in this procedure to take as this value the *mid-point* of the step. In this case the mid-point is 51. Since by the previous calculation it has been shown that the correction for this assumed mean must be made in a negative direction the

[5] The question of how many places to carry decimals constantly arises in statistical work. In general, in all work in this book calling for the calculation of the correction (c) carry the decimals to three places and round to two. For example, in the problem above the division of 18 by 37 results in a decimal of .486, but in its use as *c* in computing the mean it is rounded to .49. If the decimal were .484 the four in the third place would be dropped and the value used as *c* would be .48. In all work in this book in which square roots are to be taken of decimal values the decimals should be carried to four places and rounded to three before the root is taken. All answers to the problems in this book are computed on this basis.

value of 1.47 must be subtracted from 51, which leaves 49.53 as the arithmetic mean of this distribution calculated by means of the shorter method.

The actual arithmetic mean of the 37 scores in this frequency distribution found by the long process is 49.78. The student will do well to recheck all of the work of finding this arithmetic mean from the data in Table 9 in order to become familiar with the various steps in the computations.

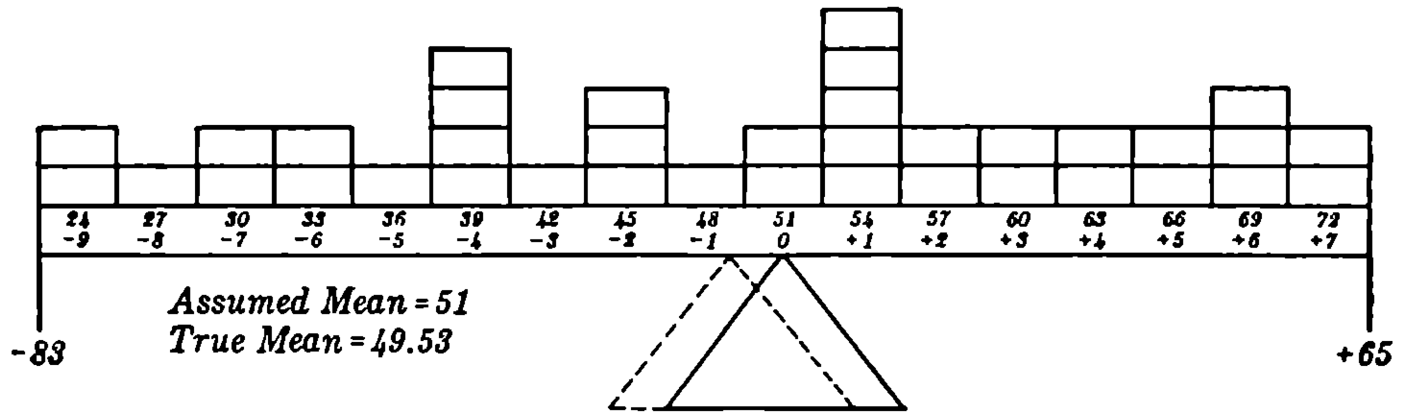


FIGURE 6. ILLUSTRATING THE CALCULATION OF THE MEAN

90. Summary of Steps in Finding Arithmetic Mean. The following is a summary of the specific steps involved in calculating the arithmetic mean from a frequency distribution by the shorter process.

1. Take the mid-point of some likely step on the scale as the *assumed mean*. Call this assumed point *zero*.
2. Lay off *steps of deviation* above and below this assumed mean maintaining the usual algebraic signs.
3. Find the *fd's* by multiplying the frequencies by the deviations. Carry the signs.
4. Find the algebraic sum of these *fd's* keeping the sign of the result.
5. Divide this resultant value by the number of cases in the distribution, and multiply this value by the number of units per step. This result is the *correction* (c).
6. In accordance with the sign of this correction (c) add to or subtract from the value assumed as the mid-point or the assumed mean of the scale (the step taken as zero). This should give the *arithmetic mean.*

91. Classroom Use of the Arithmetic Mean. The practical usefulness of the arithmetic mean in connection with the daily work of the classroom teacher is illustrated in the case of Miss L. At the end of a six weeks' period of instruction over a certain unit of general science Miss L. gave an objective informal test of the recall type which had been given

to similar classes in previous years. She wished to compare the achievement of this class with that of the classes which had taken the test previously. Some measure of central tendency was necessary. She therefore prepared a frequency distribution showing the scores earned by the 42 pupils in her class, and from this table computed the arithmetic mean of the scores. This arithmetic mean she was able to compare with the results of using this test in previous years. The frequency table and a copy of her work showing the computation of the arithmetic mean step by step is given in Table 11. The student will do well to check this work carefully and see that all calculations are correct.[6]

Table 11

DISTRIBUTION OF SCORES AND THE CALCULATION OF THE ARITHMETIC MEAN OF THESE SCORES ON A TRUE-FALSE TEST BY MISS L.'S NINTH GRADE CLASS IN GENERAL SCIENCE

Class Intervals	f	d	fd
112.5 (115) 117.5	1	+6	+6
107.5 (110) 112.5	0	+5	+0
102.5 (105) 107.5	1	+4	+4
97.5 (100) 102.5	3	+3	+9
92.5 (95) 97.5	6	+2	+12
87.5 (90) 92.5	4	+1	+4
82.5 (85) 87.5	9	0	(+35)
77.5 (80) 82.5	5	-1	-5
72.5 (75) 77.5	3	-2	-6
67.5 (70) 72.5	2	-3	-6
62.5 (65) 67.5	3	-4	-12
57.5 (60) 62.5	1	-5	-5
52.5 (55) 57.5	2	-6	-12
47.5 (50) 52.5	1	-7	-7
42.5 (45) 47.5	1	-8	-8
Total (N) = 42			(-61)

Step 1. Assumed Mean = 85
Step 2.
Step 3.
Step 4. +35 - 61 = -26
Step 5. $\frac{-26}{42} = -.619$
$-.619 \times 5 = -3.095$
Step 6. 85.000
3.095
81.905 or 81.91

My class average last year on this same test was 70.2 points, showing a slight superiority in the achievement of this year's class.

[6] A large amount of practice is needed in order to develop speedy and accurate work in computing arithmetic means. See Problems 4, 5, and 6, on pages 166 to 167 for drill on this type of work. For a large amount of material specially designed for this purpose see Greene, H. A., *Work-Book in Educational Measurements*, particularly Problems 5 to 10.

PROBLEM-EXERCISES IN COMPUTING ARITHMETIC MEANS

Problem 4

Computing the Arithmetic Mean from Frequency Distribution

a. Complete the steps in the computation of the arithmetic mean in the following frequency table.

Class Intervals	f	d	fd
147.5–152.5	1		
143–147	1		
138–142	0		
133–137	0		
128–132	1		
123–127	2		
118–122	1		
113–117	2		
108–112	3		
103–107	4		
98–102	3		
93–97	9		
88–92	9		
83–87	8		
78–82	10		
73–77	12	0	0
68–72	11		
63–67	6		
58–62	10		
53–57	8		
48–52	12		
43–47	8		
38–42	6		
32.5–37.5	3		
$N=$		$\Sigma fd=$	

Class Interval =

Assumed Mean =

Correction $=c=\frac{\Sigma fd}{N}=$

Arithmetic Mean = the Assumed Mean $+\left(\frac{\Sigma fd}{N}\times 5\right)=$

Answer (Arithmetic Mean) = 74.05

Problem 5

Computing the Arithmetic Mean.

Compute the arithmetic mean from the frequency table prepared for Problem 2, page 159. (Answer = 51.15 from frequency table with step of 3.)

Problem 6

Find the Arithmetic Mean from the tabulation prepared for Problem 3, page 160. (Answer = 31.84 from frequency table with step of 3.)

92. **The Median as a Measure of Central Tendency.** The arithmetic mean is one of the most popular and useful measures of central tendency for most research and experimental purposes. Its present popularity is undoubtedly due to the simplified methods of computing it which have been developed. However, in the interpretation of test results the median is still, and will probably always continue to be widely used.

Early workers with tests popularized the practice of taking the score of the middle paper of a pile of test papers arranged in order of size of scores as the expression of the central tendency of the group. It is the ease with which this so-called median is found that has appealed to the classroom teacher. For a long time this was called the *median*. However, more recent workers with tests have recognized the fact that the score of the middle paper in a pile of test papers stacked in order of size of scores is not the same as the middle point on the scale of a frequency table of these same scores. Accordingly a distinction is now made between the score found on the middle paper of a pile of stacked papers and the median proper. The score of the middle paper of a pile of test papers arranged in systematic order is called the *mid-measure* to distinguish it from the *median*, which is the corresponding value when the data are grouped in a frequency distribution. Thus the *mid-measure* is a counting median found from ungrouped data. The *median* is computed only from tabulated data.

93. **Computing a Counting Median or Mid-Measure.** The method of computing the mid-measure is illustrated by referring to the data given in Table 8, page 152. The al-

gebra test scores as given in Table 8 are not arranged in any systematic order. The scores from these 37 test papers are arranged in descending order in Table 12.

TABLE 12

ALGEBRA TEST SCORES GIVEN IN TABLE 8 ARRANGED IN DESCENDING ORDER

72, 71, 70, 68, 68, 66, 66, 64, 63, 60, 60, 58, 57, 55, 55, 54, 53, 53, 52, 50, 48, 46, 46, 45, 43, 40, 40, 40, 38, 36, 34, 32, 30, 30, 28, 25, 24.

The mid-measure is the score of the middle paper, or the average of the two scores nearest the middle when the number of cases is even. In this case the number of papers or scores is 37 (odd). Thus the mid-measure is 52, a score such that there are just as many equal to or larger as there are equal to or smaller than it is.[7]

94. Computing the Median from Frequency Distribution. By definition the mid-measure and the median are quite similar, the main distinction being that the *mid-measure* is designated as *an actual score on a certain paper* (or the average of the scores on two papers lying nearest the middle paper), while the *median* is defined directly in terms of *a point on the scale* of the frequency table on which it is based. *The median is a point on the scale such that 50% of the cases in the distribution are above it and 50% of the cases are below it.*

At this point it will be helpful if the student will recall that in working with the arithmetic mean from data grouped by class-intervals it was assumed that the scores falling in a given step were to be considered as having the value of the mid-point of the step. This amounts to assuming that all of the cases in a step are *concentrated at the mid-point* when computing the arithmetic mean. However, in computing the median a different assumption is made. It is assumed in

[7] Additional practice in finding the mid-measure and the median is given in Problems 7, 8, and 9, in this book and in Problems 11 to 15 of the *Work-Book in Educational Measurements,* by H. A. Greene.

this case that regardless of how few or how many cases fall in a given step these cases are *distributed uniformly through the step.* Thus, if there are three cases in a given step, one of the cases represents exactly one-third of the step. If there are 15 cases in the step one case represents one-fifteenth of the step. If this point is kept in mind much of the difficulty students find in computing the median will be removed.

TABLE 13

ALGEBRA SCORES OF A NINTH GRADE CLASS

(From Table 8)

Class Intervals	*f*	
70.5–73.5	2	$\frac{N}{2}=18.5$
67.5–70.5	3	
64.5–67.5	2	
61.5–64.5	2	Sub-total = 17
58.5–61.5	2	
55.5–58.5	2	
52.5–55.5	5	$\frac{1.5}{2.0}\times 3=2.25$
49.5–52.5	2	
46.5–49.5	1	
43.5–46.5	3	
40.5–43.5	1	
37.5–40.5	4	49.5
34.5–37.5	1	+2.25
31.5–34.5	2	51.75 = Median
28.5–31.5	2	
25.5–28.5	1	
22.5–25.5	2	
$N=37$		

The method of calculating the median is illustrated by the data given in Table 9, page 157, which is reproduced here as Table 13 in order to show all of the steps in the work. The following statements summarize the processes accurately as worked out in this illustration. To find the median of a series of scores arranged in frequency distribution :

(1) Divide the total number of measures by two to determine 50% of the cases. In this table the total number of measures (N[8]) is 37. Thus, $N/2$ equals 18.5, or the half-sum.

(2) Beginning at the bottom of the column of frequencies count up as far as possible without exceeding the half-sum, 18.5. In this case the frequencies 2 + 1 + 2 + 2 + 1 + 4 + 1 + 3 + 1 give a total of 17 without exceeding the half-sum.

(3) The difference between the *half-sum*, 18.5, and the *sub-total*, 17, shows how far into the class-interval in which the median is located it is necessary to go in order to come to the median itself. Subtracting 17 from 18.5 leaves 1.5 units which must be taken from the step in which the median is located. Since the next step contains two cases the median is located 1.5/2.0 or three-fourths of the way through the step.

(4) The two cases in the step are considered to be uniformly distributed through the interval ; hence the determination of the correction, or the fraction of the step which must be passed in counting up the frequencies to a place where 50% of the cases are found below, is merely that of taking whatever portion 1.5/2.0 is of the number of units per step. There are three units in each step in this table, therefore the correction equals 1.5/2.0 × 3, or 2.25 units.

(5) Since the correction is to be made into the interval beginning 49.5 this value 2.25 units is to be added to 49.5. The median thus becomes 51.75. Obviously, if these calculations had been made by adding the frequencies down from the top of the table the same median would have resulted, but the correction would be .75 units and would be deducted from the top of the step instead of being added to the bottom of the interval in which the correction is made. (52.5 — .75 = 51.75). This fact in itself affords a very convenient check on the calculation of medians. The student will find that it is an excellent practice to check the accuracy of all calculations in these illustrations.

Problem 7

Computing the Mid-measure.

Find the mid-measure or counting median from the data presented in Problem 1, page 159. (Answer = 33.)

Problem 8

Finding the Mid-measure.

Find the mid-measure for the data in Problem 2, page 159. (Answer = 50.)

[8] N means the total number of cases when referring to a frequency table.

Problem 9

Tabulating, Finding Median, and the Arithmetic Mean.

The following scores represent the number of physics problems worked correctly by the individuals in a class of 29 students :

4, 11, 17, 12, 10, 9, 10, 11, 9, 11, 2, 14, 12, 10, 9, 14, 13, 16, 11, 13, 14, 11, 8, 8, 11, 9, 14, 12, 10.

a. Make a frequency distribution of these scores.
b. Compute the Median of the scores in the distribution. (Answer = 10.92.)
c. Find the Arithmetic Mean of the scores in the distribution. (Answer = 10.86.)

III. Measures of Variability

95. Need for Measures of Variability. The measures of central tendency represented by the arithmetic mean and the median are valuable statistical measures but they describe only one characteristic of the data, the tendency of the scores to pile up at or near the middle of the distribution. Descriptions of test results based wholly on one or the other of these measures are incomplete.

Table 14

DATA SHOWING IDENTICAL MEANS, UNLIKE VARIABILITY

Class A		Class B
122		98
116		96
108		95
101		93
96		90
92		89
89		87
86	Mn.	86
83		85
80		83
76		82
71		79
64		77
56		76
50		74

The two groups of scores presented as Class A and Class B in Table 14 illustrate this situation very clearly. The means of the two series of scores are identical, each being 86. The *range* of the scores for Class A is 72 (122 — 50) which is exactly three times the range (98 — 74 = 24) of scores for Class B. The *Quartile Deviation* computed from the ungrouped scores is 15 for Class A and 7 for Class B. The *Standard Deviations*[9] of the scores are 20.16 and 7.3 for the Class A and Class B scores respectively. Even the most inexperienced teacher or student must recognize that very different ranges of ability are present in these two classes and that correspondingly different instructional problems are presented to the teacher.

A graphic illustration based on other data showing the unlikenesses which may appear in distributions having the same mean is given in Figure 7.

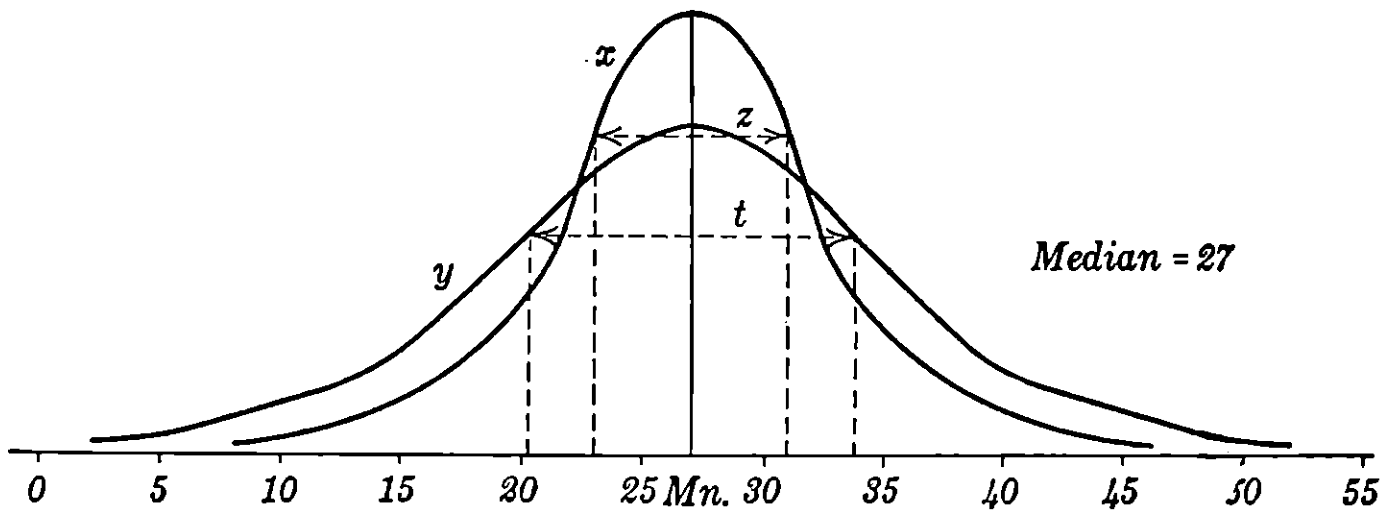


FIGURE 7. ILLUSTRATING NEED FOR MEASURES OF VARIABILITY

It is a common practice to show frequency distributions in graphic form by representing the frequencies at a given point on the scale in terms of a line erected perpendicular to the base line or scale. If the tops of a large number of these perpendiculars are connected, the result is a curved line which usually is close to the base line at the ends of the scale but rises quite rapidly from the base line near the middle of the scale. In Figure 7 the curve marked *x* represents the distribution of scores made on a certain test by a class. The curve being close to the base line at the ends or extremes means that there are relatively few very low and

[9] The quartile deviation and the standard deviation are explained in detail later in this chapter.

very high scores. The high point of the curve near the middle indicates that a great many pupils made scores near the average. This is typical of situations usually found where considerable numbers of cases are involved.

It will be noted in Figure 7 that while the middle portion of the curve *x* rises much higher than the similar portion of the curve *y* the extremes of curve *x* do not go out on the base lines in either direction as far as is true of curve *y*. This flatness or peakedness of the curve is the graphic indication of the variability of the data it represents. The less peaked the curve the greater the variability, other things being equal. It is thus apparent from this illustration that while the means of these two distributions are identical, very greatly different teaching problems are represented. Curve *x* represents a relatively homogeneous group while curve *y* is a more widely scattered group. If these two curves represent the distribution of ability of two classes, which one would you prefer to teach?

96. **Range as an Expression of Variability.** The *range* of scores, that is, the scale distance between the lowest and highest scores in a distribution is one way of expressing variability. However, it is one of the least reliable measures of variability or dispersion since it is apparent that it is affected by the fluctuaton of the extreme scores.

97. **"Q" or Quartile Deviation as a Measure of Variability.** A much more reliable expression of variation is found in the *semi-inter-quartile-range* which is really the range of the middle 50% of the cases. Thus the operation of the unusual deviation of "sport" cases at the extremes of arrays is eliminated. In actual practice *the range of the middle half of the cases is expressed in terms of one-half of the interquartile range and is called the quartile deviation, or Q.* The values of the upper quartile (Q_3) and of the lower quartile (Q_1) correspond respectively to the seventy-fifth percentile and the twenty-fifth percentile. The value Q and the median (Q_2) are distinct, however. One is a measure of variation or spread while the other is a measure of central tendency. The difference between the median (*a measure of central tendency*) and the quartile deviation (Q, *a measure of variability*) is made clear by Figure 7, page

172. The median line is the heavy vertical line labeled *Mn* which cuts the base line at the value 27 on the scale. The median is a *point* on the scale. The value Q is a *distance* on the scale, and in practice is represented by one-half the length of the lines t or z in the diagram. The difference in the variability of the two distributions represented by the curves x and y is shown by the relative lengths of the lines t and z. In practice the value Q for the curve x is stated as one-half of the line z.

It has been pointed out that the central tendencies of the groups represented by the curves x and y in Figure 7 are identical while the spread of ability is much greater in distribution y than in x. Note that line t which represents the entire range of the middle 50% of the cases in distribution y, or two times the Q for this distribution, takes in a much wider range of scores than is the case for line z which represents the entire range of the middle 50% of the scores in distribution x. All measures of variability indicate one way or another the average or typical amount of deviation from the average of the group as a whole. Central tendencies reveal the most likely point of balance between the high and low scores in a distribution. Measures of variability show how the group scatters from this central tendency. A distribution with a large measure of variability scatters more widely from the central tendency than does a distribution with a smaller measure of variability.

A comprehension of this general concept underlies an understanding of practically all of the measures of variability, as well as a great many of the statistical techniques yet to be considered in connection with the interpretation of tests.

The quartile deviation (Q) and the standard deviation, or sigma (σ) are representative of the more important measures of variability, and constitute the ones treated intensively in this book.

98. Computing the Quartile Deviation (Q). The method of computing the quartile deviation (Q) is demonstrated by the use of the data given in Table 15, page 175.

The first quartile or twenty-fifth percentile and the third quartile or the seventy-fifth percentile of a distribution are found in the same manner as is employed in finding the

median except that, of course, the numbers of cases counted off above or below the points on the scale are different. For the first quartile one-fourth or 25% of the cases are counted off below the point taken as Q_1. For the third or upper quartile 75% of the cases are counted off below the point taken as Q_3. This point (Q_3) may be more easily obtained by counting off 25% of the cases from the top of the distribution.

The following statements of the specific steps necessary in the calculation of the quartile deviation (Q) are based upon the data given in Table 15.

TABLE 15

A FREQUENCY DISTRIBUTION

Class Intervals	f
147.5–152.5	2
142.5–147.5	3
137.5–142.5	0
132.5–137.5	0
127.5–132.5	1
122.5–127.5	1
117.5–122.5	1
112.5–117.5	2
107.5–112.5	3
102.5–107.5	2
97.5–102.5	3
92.5– 97.5	9
87.5– 92.5	9
82.5– 87.5	8
77.5– 82.5	8
72.5– 77.5	12
67.5– 72.5	8
62.5– 67.5	6
57.5– 62.5	10
52.5– 57.5	8
47.5– 52.5	12
42.5– 47.5	8
37.5– 42.5	6
32.5– 37.5	3
$N = 125$	

Step 1. The third or upper quartile (Q_3) is a point on the scale of this table such that 75% of the cases are below it and 25% are above it. Therefore count into the distribution from the upper end until one-fourth ($N/4$) of the cases are taken. Make the interpolation into the class-interval in exactly the same manner as is done in finding the median. In this illustration $N/4$ is 31.25 cases. There are 27 cases down to the top of the step beginning 92.5. This leaves 4.25 to be taken from the 9 cases in this step. The resulting interpolation involves taking 4.25/9 of a step of 5 which equals 2.36. This amount must be subtracted from the *upper* limit of the step (92.5) which gives a Q_3 of 90.14.

Step 2. Q_1 is found by counting into the distribution from the lower end one-fourth or 25% of the cases. In this illustration $Q_1 = 53.91$.

Step 3. The Quartile deviation or Q is found by taking one-half of the difference between the values of Q_1 and Q_3. The formula is $Q = \frac{Q_3 - Q_1}{2}$. In this illustration

$Q = 18.11$, which means that for this distribution a distance of 18.11 units in both directions from the median will include one-half of the cases in the distribution when the distribution is approximately symmetrical or *normal.* Most distributions of educational test results are sufficiently normal to permit this interpretation.

Problem-Exercises[10]

Problem 10

Find the quartile deviation (Q) of the scores in the distributions prepared for (a) Problem 2, and (b) Problem 3. (Answer: (a) 9.19; (b) 2.78.)

99. Standard Deviation (σ, sigma) as a Measure of Variability. Such simple methods of expressing the variation of a distribution of test scores as the range and the Quartile Deviation (Q) are sufficient for most ordinary statistical situations but where careful work of an analytical or research type is being done a more exact means of expressing variability must be used. The standard deviation, also called sigma (σ), has many characteristics which make it a useful measure of variability. *The standard deviation is a sort of an arithmetic mean of the deviations from the mean of the distribution.* It is a special type of mean of the deviations because of the method used in computing it. In calculating the standard deviation (σ) each individual deviation from the mean is squared, the sum of these squared values obtained is then divided by the number of such deviations, and the square root of the result is then obtained. Restated, *the standard deviation (sigma) is the square root of the mean of the squares of the deviations from the mean of a distribution.* Expressed in symbols the standard deviation becomes $s\sqrt{\frac{\Sigma fd^2}{N} - c^2)}$ where Σfd^2 equals the deviations in the form of the sum of the products of the frequencies at each step by the deviation of each step from the assumed mean; N equals the number of cases in the dis-

[10] For further practice on the computation of Q, the student is referred to Problems 17 to 21 inclusive in the *Work-Book in Educational Measurements.* Skill in the application of these statistical methods in the interpretation of your test results will be developed and maintained only through extensive practice on practical problems.

tribution ; c equals the correction as found in calculating the mean ; and s equals the size of the class-interval of the distribution in units.

The relation of the standard deviation to the quartile deviation of a distribution is shown in the accompanying figure (Figure 8). By definition the *quartile deviation* (Q) takes into account the middle fifty per cent of the cases. That is, the ordinates (lines erected perpendicular to the base line of the curve) erected at a scale distance equal to Q on either side of the mean or median include fifty per cent of the area of the surface between the curve and the base line. In the diagram the lines P and R represent the ordinates erected at a distance equal to Q on either side of the mean. The lines S and T represent ordinates erected at a distance equal to σ on either side of the mean. The standard deviation (σ) takes into account approximately 68 per cent (in a normal distribution 68.26%) of the area of such

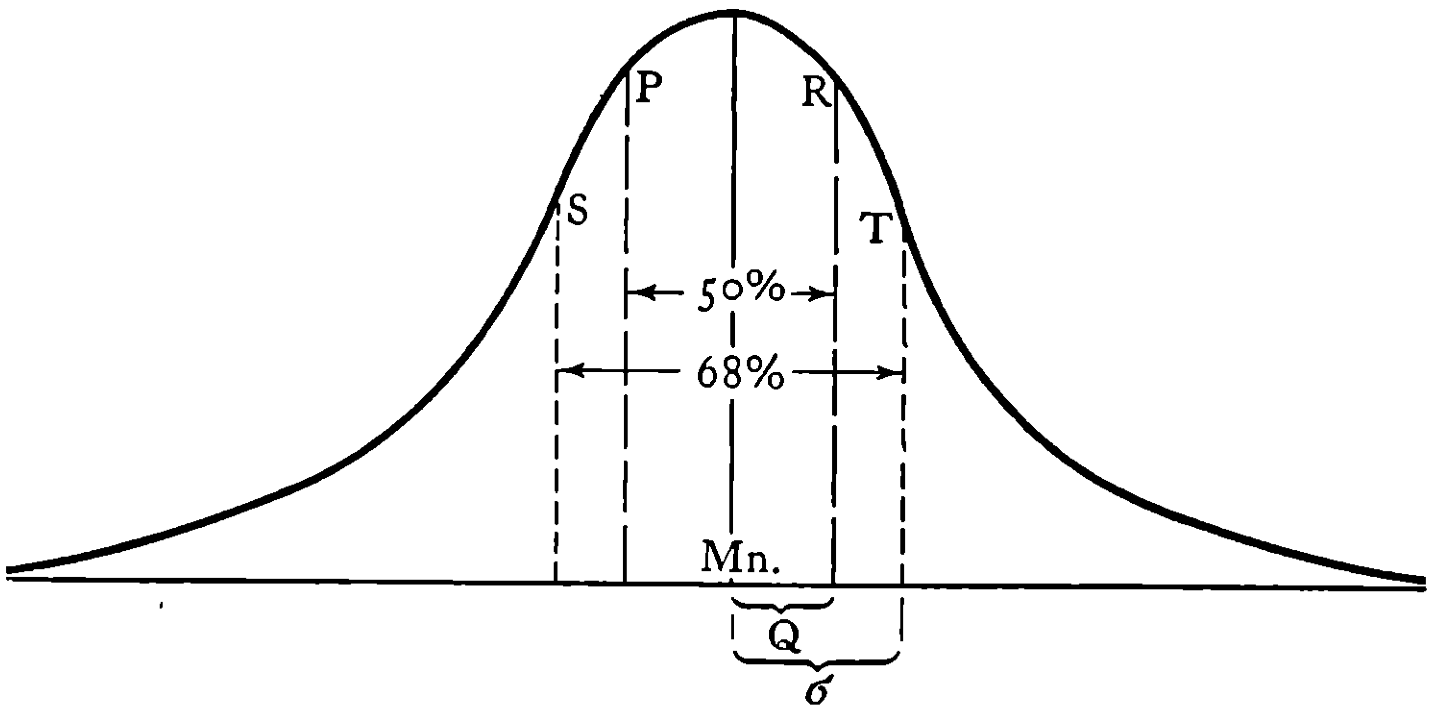


FIGURE 8. COMPARISON OF STANDARD DEVIATION (σ) AND QUARTILE DEVIATION (Q)

a distribution. That is, ordinates erected at a distance equal to sigma on either side of the arithmetic mean include approximately two-thirds of the cases in the distribution.

The standard deviation (σ) may be computed about any measure of central tendency but it is common practice to consider it as computed about the arithmetic mean only. Unless otherwise indicated sigma (σ) may be considered as having been calculated about the arithmetic mean.

A further interesting characteristic of the standard deviation is indicated in Figure 8. Mathematically the value

sigma bears a definite relationship to the curve of the distribution itself. It will be noted that in a distribution involving any large number of cases or scores there is a tendency for the larger portion of the cases to pile up in or near the middle of the distribution. When a normal distribution is presented in graphic form the result is a symmetrical bell-shaped curve with many cases in the middle and few cases at the extremes. Certain types of these characteristic bell-shaped distributions have come to be called *normal curves*. For these normal curves, formulae have been derived from which such typical curves may be computed if certain basic data concerning the curve are given. In these formulae *sigma* is one of the values which must be given in order to construct such a curve. Sigma, in the typical formula, represents the distance from the mean at which the curve changes from convex to concave. In Figure 8 the points where the curve changes its character are indicated by the ordinates lettered *S* and *T*.

Thus, because of this direct mathematical relationship which the standard deviation bears to the curve of the distribution itself, and the reliable expression of variability which it provides due to the fact that every deviation in the distribution is considered, the standard deviation is one of the most useful of the measures of variability.

100. Computing the Standard Deviation (σ) from Ungrouped Data. In the computation of the standard deviation from ungrouped data, as in the accompanying illustration, the mean of the distribution must be found. When the data are grouped in a frequency table it is not strictly necessary for the arithmetic mean to be computed, although it is necessary to go through all but the last step of the process.

The steps in the computation of the standard deviation from ungrouped data are given in detail (Table 16) in connection with data from Table 14 on page 171. The mean of the scores for Class A in Table 14 is 86. Thus, a score of 89 deviates from this mean by 3 points. A score of 96 deviates 10 points, etc. (See column *d* in the table.) The standard deviation (σ) is the square root of the mean of the squares of these deviates from the mean of the array of

scores. It is necessary, therefore, to square each of these deviates. These are given under the column headed d^2. Since each deviation appears only once and the data are ungrouped the formula may be simplified to read $\sigma = \sqrt{\frac{\Sigma d^2}{N}}$. The sum of the deviation squared (Σd^2) in this case is 6100. The mean of these deviations is therefore 40.667. This value is the mean of the deviation squared. Therefore, to turn it into units of the scale, the square root of this quantity must be taken. This value is 6.38, which is the standard deviation (σ) of this series of scores. The mean of this distribution is 86. The σ is 6.38. This means that between scores 6.38 points larger and 6.38 points smaller than this mean, approximately two-thirds (68.26%) of the scores will be found.

TABLE 16

STANDARD DEVIATION FROM UNGROUPED DATA

Data for Class A from Table 14

Test Scores	*d* (Deviations)	d^2 (Deviations Squared)	
122	+36	1296	
116	+30	900	
108	+22	484	
101	+15	225	
96	+10	100	$\sigma = \sqrt{\frac{\Sigma d^2}{N}}$
92	+ 6	36	
89	+ 3	9	
86	0	0	$= \sqrt{\frac{6100}{15}}$
83	− 3	9	
80	− 6	36	
76	−10	100	$= \sqrt{40.67}$
71	−15	225	
64	−22	484	$= 6.38$
56	−30	900	
50	−36	1296	
Total = 1290 Mean = 86		$\Sigma d^2 = 6100$	

101. **Computation of Sigma from Grouped Data.** The method of computing the standard deviation from ungrouped data illustrated in Table 16 may be applied with very few changes to the calculation of sigma from grouped data. A slight change in the general formula is required, for when the scores are grouped in class intervals, the deviations of the scores must be considered by groups having the mid-points of the steps in which they are found. This permits the expression of the deviations in steps rather than in units of the scale. The formula for use in calculating the standard deviation when the data are grouped in a frequency distribution is $s\sqrt{\frac{\Sigma fd^2}{N}-c^2}$. The steps in the application of this formula in the calculation of the standard deviation of the scores presented in Table 16 will make clear all of the processes involved in finding the sigma of a frequency distribution. The computations themselves are shown in Table 17.

Step 1. Assume a mean as near as possible to the true mean of the distribution in order that the correction (c) may be as small as possible. In case the correction is larger than .5 steps it may be desirable to start the work over and assume a mean nearer the true mean. In Table 17 the step with a mid-point of 7 is taken as the zero step.

Step 2. Lay off deviations above and below the step assumed as zero and multiply the frequencies in each step by the deviation of the step exactly as in the calculation of the arithmetic mean.

Step 3. Summate the plus *fd's* and the minus *fd's* algebraically. The sum of the *fd's* in this problem is + 2 units. Divide this by the number of cases in the table (N = 27) and the resulting correction (c) is + .074. This correction is the same as the one used in computing the mean except that it is turned into units of the scale. In computing the standard deviation it is not strictly necessary to calculate the mean of the distribution.

Step 4. Square this correction in order to have it in the same denomination as the values from which it must later be subtracted. The square of c (+ .074) is .005 in this problem.

Step 5. Multiply the values under the column headed *fd* by the values under *d*. This will give a column of values known as fd^2. Summate this column. In this problem the fd^2 is 218.

Step 6. Divide the sum of the fd^2 by N, the number of cases in the distribution. The result of this division is 8.037.[11]

Step 7. Since this value, 8.037, is always too large in proportion to the amount the true mean deviates from the assumed mean (in this case, the amount represented by the value *c*) it must be reduced an amount equal to the square of *c*. Thus, 8.037 — .005 = 8.032.

Step 8. The sigma of this distribution expressed in steps is next obtained by extracting the square root[12] of the value 8.032. The square root of this value to two decimal places is 2.83. Since the class-intervals used in this frequency table are steps of one unit, the standard deviation is therefore 2.83.

This means that if ordinates are erected at a distance of 2.83 units on either side of the arithmetic mean of this distribution approximately 68% of the distribution will be found between these ordinates. Of course this will not be strictly true, since no small frequency table such as the one used in this illustration will be sufficiently symmetrical to permit the cases to fall in exactly that fashion. Roughly, this value, sigma, expresses the range of the middle 68% of the cases.

The arithmetic mean of the data used in this illustration is 7.07. The standard deviation (σ) in points on the scale is 2.83. Ordinates erected at 2.83 points above this arithmetic mean, and at 2.83 points below it fall at 9.9 and 4.24

[11] This value prior to the application of the correction, c^2, is frequently designated as the *root-mean-square-deviation* or S^2. The correction, *c* (squared to make it comparable to the S^2) is always deducted from the root-mean-square-deviation because this value, S^2, is always too large. If *c* has any value more than zero it means that the assumed mean and the true arithmetic mean do not coincide; therefore, any deviations computed about it are too large and the correction *c* must be subtracted to compensate for this difference.

[12] The student will find the tables of squares and square roots, such as are given on pages 303-312 of Garrett's *Statistics in Psychology and Education* or Otis' *Statistical Methods in Educational Measurements*, pages 299 to 305, very helpful in his work from this point on. Ability to use a slide rule or a mechanical calculator will also speed up the work as well as make it more accurate.

TABLE 17

STANDARD DEVIATION FROM FREQUENCY DISTRIBUTION

Class Intervals	Mid-Points	f	d	fd	fd^2
14.5–15.5	15	1	+8	8	64
13.5–14.5	14	0	+7	0	
12.5–13.5	13	0	+6	0	
11.5–12.5	12	1	+5	5	25
10.5–11.5	11	1	+4	4	16
9.5–10.5	10	2	+3	6	18
8.5– 9.5	9	2	+2	4	8
7.5– 8.5	8	3	+1	3	3
6.5– 7.5	7	4	0	+30	
5.5– 6.5	6	6	−1	6	6
4.5– 5.5	5	3	−2	6	12
3.5– 4.5	4	1	−3	3	9
2.5– 3.5	3	2	−4	8	32
1.5– 2.5	2	1	−5	5	25
		$N=27$		−28	218

$$c=\frac{30-28}{27}=\frac{+2}{27}=+.074$$

$$c^2=.005$$

$$\frac{\Sigma fd^2}{N}=\frac{218}{27}=8.037$$

$$8.037-.005=8.032$$

$$\sigma\text{steps}=8.032=2.83$$

Step of one in table, therefore $\sigma=2.83$[13]

(Arithmetic Mean $=7.07$ or 7.1)

on the test scale. Within these points will be found approximately 68% of the measures in the distribution when it represents enough cases to be approximately symmetrical.

102. **Practical Uses of the Standard Deviation.** *Assign-*

[13] If the step used in Table 17 were one or three or five, the value of sigma in steps (2.83) would be multiplied by the size of the step, three or five, as the case might be.

ment of Class Grades. The student or teacher who is interested in the critical analysis of test scores will find the standard deviation a very useful and reliable instrument for the purpose. For example, it offers the basis for an objective plan for turning scores on objective tests into class marks. The importance of this practice is so great that the steps involved in the technique are given in detail. The computations described are based upon the objective test scores from a class of 45 pupils given in Table 18. The student will do well to check all of these computations to see that they are free from errors.

Step 1. Prepare a suitable frequency table of the test scores, lay off the deviations from the assumed mean, find the sum of the *fd's*, and the arithmetic mean. The mean of this distribution is 68.55.

Step 2. Compute the standard deviation (σ) of this distribution in exactly the same way as illustrated on page 182. Be sure to study each computation given in Table 18.

Step 3. Since a distance of two and one-half sigma units above and below the mean includes almost 99% of all cases in a distribution, this number of sigma units is laid off above and below the mean to mark off the limits for the five marks of A, B, C, D, and Fd. This naturally results in placing one of the sigma-units in the middle of the distribution in such a way that one-half of the sigma distance of the middle unit extends above the mean and one-half below. Accordingly, to the arithmetic mean of 68.55 add one-half of the standard deviation (one-half of 19.40). This gives a value of 78.25 which becomes the upper limit of the group of scores which will be assigned grades of *C*.

Step 4. The upper limit of the group of scores to be assigned *B's* is found by adding one and one-half standard deviation units to the arithmetic mean. Thus, $68.55 + 1.5 (19.40) = 97.65$, which is the upper limit of the B group.

Step 5. The upper limit of the D group is found by subtracting one-half of a standard deviation unit from the mean. ($68.55 - .5 (19.40) = 58.85$.)

Step 6. The upper limit of the *Fd* group is obtained by subtracting one and one-half sigma units from the mean of the distribution. ($68.55 - 1.5 (19.40) = 39.45$.)

Table 18

STANDARD DEVIATION TECHNIQUE FOR ASSIGNING CLASS GRADES[14]

Test Scores	Grade
109, 104, 103, 102, 99	A (11.1%)
95, 95, 94, 93, 84, 83, 79, 79	B (17.8%)
77, 76, 76, 71, 71, 69, 64, 64, 64, 64, 62, 60, 60, 60, 59, 59	C (35.5%)
58, 57, 57, 56, 56, 55, 55, 53, 52, 52, 51, 51, 47, 41	D(31.1%)
37, 37	Fd. (4.5%)

Mid-Points	Class Intervals	*f*	*d*	*fd*	*fd*²
110	107.5–112.5	1	10	10	100
105	102.5–107.5	2	9	18	162
100	97.5–102.5	2	8	16	128
95	92.5– 97.5	4	7	28	196
90	87.5– 92.5	0	6	0	0
85	82.5– 87.5	2	5	10	50
80	77.5– 82.5	2	4	8	32
75	72.5– 77.5	3	3	9	27
70	67.5– 72.5	3	2	6	12
65	62.5– 67.5	4	1	4	4
60	57.5– 62.5	7		(+109)	
55	52.5– 57.5	7	−1	−7	7
50	47.5– 52.5	4	−2	−8	16
45	42.5– 47.5	1	−3	−3	9
40	37.5– 42.5	1	−4	−4	16
35	32.5– 37.5	2	−5	−10	50
	N =	45		(−32)77	809

$$A.M. = 60 + s\,\frac{\Sigma fd}{n} \qquad S.D. = s\sqrt{\frac{\Sigma fd^2}{n} - c^2}$$

$$= 60 + 5 \cdot \frac{77}{45} \qquad = 5\sqrt{\frac{809}{45} - (1.71)^2}$$

$$= 60 + 5\,(1.71) \qquad = 5\sqrt{17.98 - 2.92}$$

$$= 60 + 8.55 \qquad = 5\sqrt{15.06}$$

$$= 68.55 \qquad = 5 \times 3.88 \text{ or } 19.40$$

Find score limits:

68.55+ ½ (19.40) = 78.25 upper limit of C group
68.55+1½ (19.40) = 97.65 upper limit of B group
68.55− ½ (19.40) = 58.85 upper limit of D group
68.55−1½ (19.40) = 39.45 upper limit of F group

A = Above 97.65 *D* = 39.45 to 58.85
B = 78.25 to 97.65 *Fd* = Below 39.45
C = 58.85 to 78.25

[14] For a complete explanation and discussion of the many problems involved in objectifying teachers' marks see Bangs, C. W., and Greene, H. A., *Teachers' Marks*

Step 7. From these values the score limits of this distribution may be set up. Class grades may then be assigned as indicated to the scores within the limits specified.

Grades	Score Limits
A	97.65 and above
B	78.25 to 97.65
C	58.85 to 78.25
D	39.45 to 58.85
Fd.	Below 39.45

It is readily apparent that practically no subjective factors are involved in the assignment of grades by this method. The objective test scores of the forty-five pupils used in the illustration are changed by this treatment into 5 A's, 8 B's, 16 C's, 14 D's, and 2 Fd's. he score limits are determined by the standard deviation units and would be the same no matter who assigned the grades. It should be noted, however, that these limits hold only for this particular distribution and must not be assumed to be true for any other test. The teacher should also remember that this method of grading does not take into account the absolute level of ability at which a particular class works. The superior pupil in an average or poor class receives an A by this method just as readily as does the superior pupil in a very superior class. This is probably less serious than it sounds, however, for most class groups large enough to warrant the application of this technique average out quite well in this respect.

Basis for T-Scores and other derived scores. The standard deviation also affords the basis for derivation of a number of useful derived-scores in test interpretation. For example, the T-score now commonly used in commensurating test scores depends upon the standard deviation for its basic unit. For many years prior to the development and popularization of the T-score, test scores were expressed in terms of their position in the total distribution. For instance, a pupil's score might be a number of a distribution having a mean of 60 points and standard deviation of 5 points. A score of 50 in the test would be designated by a

and the Marking System, University of Iowa Extension Bulletin No. 244, College of Education Series No. 26, May 15, 1930.

standard score of —2.0 sigma units, since it lies exactly two standard deviation units below the mean of the distribution. This same procedure is used in assigning T-scores. The formula for the T-score is $T = \frac{10(m - x)}{\sigma} + 50$ in which m is the mean of the distribution, x the deviation of the score and σ the standard deviation of the distribution. The difference between the mean and any test score is multiplied by ten in order to remove all decimal points. The fifty points are added in order that there may be no negative scores. The T-score is a very convenient method of interpreting test scores. T-scores of 25, 50 and 75 mean that the individual pupil's scores were exactly at the lower quartile, the median and the upper quartiles. This fact makes it easy to attach meaning to the test scores.

Scaling of test items. The standard deviation, along with certain other measures of variability, represents a convenient unit in which to evaluate the difficulty of test items. When used under these conditions the standard deviation of a theoretically normal curve of the specified ability is used as the unit in laying off differences in difficulty along a linear scale. As a first step in the procedure, the percentage of pupils failing on each item or exercise must be secured. By means of tables based upon the normal curve these percentages of failure are changed into standard deviation units which express the positions of each of the exercises with respect to the mean ability of an infinite and normal population. Exercises which are answered successfully by 50% of the class are assigned a position at the mean. Exercises missed by 55 or 60 per cent of the class are given sigma values above the mean, etc. A significant feature of this procedure, however, is the fact that a difference in difficulty of five per cent near the mean results in a relatively small sigma difference, while a five per cent difference near the extremes of the distribution makes a relatively large sigma difference. This is in conformity with the fact that because of the height of the curve near the mean a smaller distance along the linear scale on the base line is required to add a given area of the curve. Thus, the difference in the sigma values assigned to two test items having percentages of failure

of 55% and 60% is .13 standard deviation units, (2.74 — 2.61) ;[15] while the difference in apparent difficulty of two items failed by 90% and 95% of an experimental group is .34 standard deviation units (4.09 — 3.75). The net result of this method of item evaluation is to magnify somewhat the simplicity of the very easy item and the difficulty of the very hard one.

Sigma-units are also utilized in the construction of scales for the estimation of the merit or quality of certain classroom products. The use of these units in the derivation of such scales is too limited to warrant treatment in this book.

PROBLEM-EXERCISES

Problem 11

Compute the standard deviation of the scores in the distributions prepared for (a) Problem 2 and (b) Problem 3. (Answers: Problem 2, 13.2 ; Problem 3, 8.37.)

Problem 12

Utilize the results of your work on Problem 11 and assign T-score equivalents to the test scores used in Problem 2.

Problem 13

Assume that the language test scores used in Problem 3 are objective examination scores which you are to turn into letter grades, A, B, C, D, and Fd. using the standard deviation technique outlined and illustrated in this chapter.

IV. THE RELATIONSHIP OF TEST SCORES

103. Need for Measures of Relationship. The discussion up to this point in this chapter has concerned itself with two types of descriptions of test scores; namely, (1) measures of central tendency, and (2) measures of variability. In connection with the more critical examination and interpretation of test results it often becomes necessary for the teacher or student to go somewhat beyond the description afforded by these simple measures. It is often necessary to

[15] See Table 5 on page 392 of Rugg *Statistical Methods Applied to Education* (Houghton Mifflin) or any other similar table of sigma values.

select for testing purposes the one test of a series which most nearly measures the desired ability. The method followed in such a case involves finding the relationship between the several tests under consideration and the ability to be measured. This is called the method of *correlation*. It is the method applied when two or more measures of the same individuals are being studied with the view to finding the factors, elements or characteristics which are common to each. Sometimes the method of correlation is mistakenly used to attempt to discover causes operating to produce certain effects. There is nothing in the method or the result of computing a correlation which indicates definitely which of the factors is a cause and which is an effect, or whether both of the factors may be affected by other variables.

104. **The Coefficient of Correlation.** In the expression of relationships, as in the other statistical measures, it is desirable that this relation between two series of variables be expressed in a single objective or mathematical value. A number of different ways have been proposed for the derivation of these expressions of relationship but no one of these methods has produced a term which is both objective and easily interpreted. Methods of computing relationships in terms of the correspondence between rank positions of scores, and in terms of the percentage of the scores falling within a specified unit of variability of each other have been proposed. In general these procedures lack sufficient exactness to warrant their common use in the analysis of test results. The student who is interested in these different methods of revealing relationships will find them discussed in certain of the treatments on statistical methods listed in the references at the end of this chapter.

It is probably useless for the student of educational measurements to attempt to master more than one method of computing the coefficient of correlation. The one method considered here, the *Pearson Product-Moment method*, is by far the most common and on the whole the basic method used in educational investigations. This method, while somewhat complicated and difficult because of the large number of different calculations to be made, really involves very little that the student has not already mastered.

The Pearson Product-Moment coefficient of correlation indicated by r *is a single numerical index which expresses the extent to which the pairs of corresponding measures of two variables tend to deviate similarly from their respective arithmetic means.* The values of *r* may vary all the way from + 1.0 indicating perfect positive relationship, through all of the possible decimals to zero (0) indicating no relationship whatever, to — 1.0 indicating a perfect negative relationship. The following are illustrations of a positive relationship between two factors:

(1) The rise and fall of a column of mercury in a thermometer with the rise and fall of the outside temperature. As the temperature rises the column of mercury also rises.
(2) The amount of the Sunday School collection and the number of pupils in attendance.
(3) The direction of the wind and the movement of smoke from a chimney. The smoke moves away with the wind.
(4) The tendency of pupils who are intelligent to be good silent readers.

Negative correlation may be illustrated by:

(1) The movement of the elevator cage and the counter-balancing weights. As the elevator cage goes up the counter-balancing weights move in the opposite direction.
(2) The relation between absence from school and school achievement.

Zero or indifferent correlation is best illustrated by means of the chance matching of numbered cards which have been shuffled. Two packs of 25 blank cards each may be numbered and the packs carefully shuffled so that the cards are in no definite order. If cards are drawn from each pack and paired the resulting relationship is zero, or indifferent. If these same packs of cards are both arranged in ascending order and the first card from one is paired with the first card from the other the resulting relationship is positive. If one pack is inverted and each time a small-numbered card is taken from the one pack a large-numbered card is taken from the other the resulting correlation is negative.

The illustration of the numbered cards used here suggests one of the simple methods of expressing correlation; viz., the method of ranking. If pupils are given two tests and the scores from the tests tend to place the same pupils in

the same relative positions in each series there is an indication of a positive correlation between the two tests. For example, the accompanying pairs of scores (Table 19) earned by the nine pupils listed indicate a high positive relationship between the two tests because the pupil making the highest score on Test *A* also made the highest score on Test *B*, and each other pupil in the list maintained his relative

TABLE 19

PAIRS OF TEST SCORES

Pupil Number	Score on	
	Test *A*	Test *B*
1	89	32
2	85	31
3	83	29
4	80	28
5	76	26
6	70	24
7	65	21
8	61	20
9	54	18

position. This suggests that the two tests measure abilities which have a great many factors in common. On the other hand, if it had happened that Pupil No. 1 who made a score of 89 on Test *A* had made a score of 18 on Test *B*, and Pupil No. 2 who made a score of 85 had made a score of 20 on Test *B*, etc., the resulting correlation would have been negative, and would have suggested that the two tests were inversely related. That is, the negative correlation would indicate that the presence of ability on Test *A* implied the absence of ability on Test *B*.

The coefficient of correlation is computed from data arranged in a frequency table, but the table used is a slightly different form from that employed in any of the problem work in this book thus far. The method of tabulating paired data is explained in the following section.

105. Tabulating Test Scores in Double Entry or Correlation Tables. The double entry or correlation table is used in computing a coefficient of correlation by the Pearson Product-Moment method. It is also useful frequently when more than one test is given to the same group of subjects. The principles used in setting up the class-intervals in this double entry table are identical with those used in making a simple frequency table. In fact, the correlation table may be thought of as a simple frequency table on a vertical or y-axis which also has been extended in a horizontal direction to show the place of each frequency in the distribution on the x-axis as well. Thus each tabulation or check mark in a correlation table stands for the result of two measures of the same individual. Table 21 on page 194 is an illustration of this type of double entry or correlation table. This table is also used to illustrate the steps in the computation of the coefficient of correlation itself. Table 21 shows that one pupil who read with a comprehension score of between 7.5 and 10.5 also read with a rate score of between 59.5 and 66.5 as measured by this particular test, etc. The tendency of the frequencies to group themselves along the diagonal of the table is itself an indication of the correlation which exists between the two factors.

106. Steps in the Computation of the Pearson Product-Moment Coefficient of Correlation. The calculation of the Pearson Product-Moment coefficient of correlation involves the solution of the following formula: $r = \dfrac{\dfrac{\Sigma xy}{N} - c_x c_y}{\sigma_x \sigma_y}$ in which r is the coefficient of correlation, N the number of cases, σ_x the Standard Deviation (in steps) of the distribution on the X-axis, σ_y the Standard Deviation (in steps) of the distribution of the Y-axis, c_x the correction on the X-axis, c_y the correction on the Y-axis, and Σxy the sum of the products of the deviations of each measure from the central tendency of the X- and the Y-axes. The solution of the formula involves the following specific steps:

(1) Tabulation of paired scores in a double entry table.

(2) The computation of the Standard Deviation (σ) of each axis or the distribution, leaving σ in terms of steps.

(3) The computation of the sum of the products of the deviations of each measure from the central tendency of both axes (known in the formula as Σxy).
(4) The division of the Σxy by the number of cases.
(5) The subtraction of the algebraic product of the corrections of each axis.
(6) The changing of the resulting fractional value into decimal form retaining its sign. If the sign is negative the relationship is inverse and if positive it is direct. The possible values of this decimal form range from plus 1.0 through the possible decimals to 0 and minus 1.0.

Of these steps, Numbers 1 and 2 have been explained and illustrated in previous discussion. The major new step is the determination of the sum of the *xy* products.

The name "product-moment method" implies the significant feature of the process. The relationship itself takes into account the operation of forces (frequencies) at varying distances (deviation in steps) from the point of rotation (the mean). Since each measure assumes a position with regard to both axes the resulting moments[16] must take into account this fact. For example, if a child makes a high score on each of two tests tabulated in double entry form his position in the table will be well up in the upper right quadrant of the table.[17] The moment represented by his position is, however, the product of the distance his score lies from the mean on the X-axis and the distance his score lies from the mean on the Y-axis as well. In the illustration (Table 20) such a score might be shown by a frequency in the compartment made by the intersection of the deviations plus $4y$ and plus $3x$. The resulting xy product (moment) of such a case is $4 \times 3 \times 1$ or 12. If two cases were found at that point the moment would be 24, etc. The moments or products in the upper right hand and lower left hand quadrants are always plus if the tabulation is made upward and toward the right.

[16] The reader will recall from his high school physics that the term *moment* is given to the force which tends to produce rotation in an unbalanced body. For example, a beam resting on a fulcrum is brought into balance when the moments of the forces operating on it are equal. A weight or a force of two units operating on a beam at a distance of one unit from the point where it is suspended is equalled in moment by a weight or force of one unit at a distance of two units from the point of suspension. Similarly in mathematics forces are brought into balance by equating their moments.

[17] Assuming tabulation upward and to the right.

The products in the upper left hand and lower right hand quadrants are similarly always negative in sign. The $\Sigma\ xy$ is the algebraic sum of the plus and minus products. The sign of this product determines the sign of the correlation. The complete process of computing a product-moment correlation from a double-entry table is shown in Table 21, page 194.

107. **Illustration of the Computation of the Pearson Correlation Coefficient.** The data for this illustration represent the Comprehension and Rate scores of 76 pupils as measured by a certain silent reading test (Table 21). The rate scores (the speed of reading) are tabulated along the horizontal (X) axis, and the comprehension scores are tabulated along the vertical (Y) axis. The student should study this illustration until he becomes familiar with the steps involved. Only a very few new processes are introduced. Attention is called to the fact that the columns headed d, fd, fd^2 are required in finding the standard deviations and have nothing to do with the values marked xy.

TABLE 20

ILLUSTRATING COMPUTATION OF XY PRODUCTS

X AXIS

Y AXIS

							1		4
									3
		–					+		2
									1
									0
									–1
									–2
		+					–		–3
									–4
–4	–3	–2	–1	0	1	2	3	4	

TABLE 21

CORRELATION TABLE SHOWING RELATION OF RATE AND COMPREHENSION AS MEASURED BY A CERTAIN READING TEST

Comprehension—Y	Rate—X	66.5 59.5	73.5 66.5	80.5 73.5	87.5 80.5	94.5 87.5	101.5 94.5	108.5 101.5	115.5 108.5	122.5 115.5	f	d	fd	fd^2	xy
	52.5–55.5									2	2	+7	+14	98	42
	49.5–52.5										0	+6	0		
	46.5–49.5									4	4	+5	+20	100	60
	43.5–46.5								1	4	5	+4	+20	80	56
	40.5–43.5							4	1	2	7	+3	+21	63	36
	37.5–40.5							4	1	3	8	+2	+16	32	30
	34.5–37.5						5	2	1		8	+1	+8	8	4
	31.5–34.5						5	1		1	7	0	(+99)		
	28.5–31.5				4		3				7	−1	7	7	8
	25.5–28.5			2	4		1	1			8	−2	16	32	26
	22.5–25.5			4	5						9	−3	27	81	66
	19.5–22.5		3	1	1		1				6	−4	24	96	68
	16.5–19.5		1		1		1				3	−5	15	75	30
	13.5–16.5			1							1	−6	6	36	18
	10.5–13.5										0	−7	0	0	
	7.5–10.5	1									1	−8	8	64	40
	Total (f)	1	4	8	15	0	16	12	4	16	(76)		(−103)	(772)	(+484)
	d	−5	−4	−3	−2	−1	0	+1	+2	+3					
	fd	−5	−16	−24	−30	(−75)		+12	+8	+48	(+68)				
	fd^2	25	64	72	60			12	16	144	(393)				

$$\frac{\Sigma XY}{N} = +\frac{484}{76} = +6.37$$

$$c_x = \frac{-75+68}{N} = -.092 \qquad c_y = \frac{-103+99}{N} = -.053 \qquad c_x c_y = .005$$

$$S^2_x = \frac{393}{N} = 5.17 \qquad S^2_y = \frac{772}{N} = 10.16 \qquad \sigma_x \sigma_y = 7.22$$

$$\sigma_y = 3.18$$

$$\sigma_x = 2.27$$

$$r = \frac{\frac{\Sigma XY}{N} - c_x c_y}{\sigma_x \sigma_y} = \frac{6.37 - .005}{2.27 \times 3.18} = +.882$$

There are possibly a few points in the computation of the correlation coefficient which deserve a little further comment. The work of finding the standard deviations (sigma) of each axis of the double entry table should give no difficulty. *It will be noted, however, that the sigmas in this illustration are not expressed in units of the scale but are left in terms of steps.* The reason for this lies in the fact that if the sigmas are turned into units of the scale each of the other expressions of deviation (the *xy*'s) must also be turned into similar form. This would involve a considerable multiplication with a corresponding amount of division later, all of which is useless and only introduces additional opportunities for errors in computation. The value of the coefficient of correlation as a rule is not seriously affected by the size of the step taken in making the table. It is assumed, of course, that the class-intervals will be reasonable, but it is not necessary that they be the same size on both axes.

Care must be taken to keep the sign of the corrections on each axis. If one correction is positive and the other negative the resulting $c_x c_y$ product is of course negative. This results in an algebraic subtraction of a negative quantity which serves to increase the numerator of the formula for r and accordingly to increase r itself.

The only other point which need cause the student any trouble is in determining the sum of the *xy* products (Σxy). Let us refer again to Table 21, page 194. The frequency 2 which appears in the intersections of the class-intervals 52.5 to 55.5 (*Y*) and 115.5 to 122.5 (*X*) lies above the arithmetic mean of the *Y*-axis seven (7) steps. At the same time these cases lie to the right of (above) the mean of the *X*-axis three (3) steps. The actual moment of these two cases is therefore the product of the distance they deviate from the mean of the one axis by the distance they deviate from the mean of the other axis ($2 \times +7 \times +3$ or $+42$). The sign of the deviation involved must be carefully noted. Let us take the 8 cases which lie in the class-interval 25.5 to 28.5 on the *Y*-axis. Two of these cases deviate -2 steps from the mean of the one axis and -3 steps from the mean of the other. The product-moment of these two cases is therefore $2 \times -2 \times -3$ or $+12$. The product-moment of the

4 cases is $4 \times -2 \times -2$ or $+16$. The one case which deviates -2 steps from the mean of the one axis and 1 from the mean of the X-axis has a product-moment of $1 \times -2 \times 1$ or -2. The total of the plus moments (12 and 16) is $+28$. The resultant sum of the xy's for this particular row is therefore $28 - 2$ or $+26$. The other xy's in this illustration are found in a similar manner.

From this point in the calculation of the correlation coefficient the work is merely that of substituting values in the formula itself. The sum of the xy products ($+484$) is to be divided by the number of cases in the scatter diagram (76). The result is $+6.37$. From this must be subtracted the product of the corrections[18] of the X- and Y-axes ($c_x c_y$). This reduces the numerator of the fraction represented by the formula for r to $+6.365$. The product of the sigmas of the X- and Y-axes is 7.22. The decimal represented by the fraction $\frac{6.365}{7.22}$ is $+0.882$, the coefficient of correlation (r) of the data in this illustration. In general, this coefficient of correlation indicates that there is a positive and rather significant relationship between the rate of reading of the 76 pupils tested and the comprehension with which they read.

108. **Meaning of the Correlation Coefficient.** The method of calculating the correlation coefficient as outlined in the foregoing pages is quite mechanical and, as such, can be mastered readily by most students. The interpretation of the meaning or significance of a correlation coefficient is often quite another matter, for no entirely satisfactory mechanical device for accomplishing this has thus far been developed. A number of suggestions have been made recently, however, by means of which the student may be aided in attaching meaning to the correlation coefficient.

It will be remembered that correlation is usually indicated by means of what is called a *double-entry table*, or *correlation table.* The distribution of scores in such a table is known as a "scatter diagram." The appearance of the scatter diagram itself gives some indication of the amount of relationship which exists between the two variables shown.

[18] This is the same correction used in finding the arithmetic mean.

Assuming that the scatter diagram is made by tabulating upward and toward the right, which is the best uniform practice, a high r is usually found where there is a very definite clustering of the cases along what would be the diagonal of the table. This means that the cases tend to be grouped somewhat systematically in a practically straight line running from the lower left-hand corner of the table to the upper right-hand corner of the table. This type of grouping is shown in Table 21 on page 194. As the cases scatter from the line of this diagonal, the correlation is reduced. If the cases are scattered over the table in a generally circular arrangement the resulting correlation will approximate zero. As the relationship changes from positive to negative the elliptical grouping of the cases takes place along a diagonal running from upper left to lower right. After some experience with "scatter diagrams" the student will soon come to have a definite feeling with regard to the probable magnitude of the correlation to expect from a certain type of distribution.

One of the important outcomes of the use of correlation methods is that within certain limits it makes possible the estimating of unknown values from known values. The accuracy of this estimate, however, depends directly upon the correlation between the factors measured. For example, if it is known from previous experience that there is a high positive relationship between achievement in a specific subject and the student's response to a selected aptitude test, the probable achievement of a group of students in this course may be determined within limits by securing their response to the aptitude test. A correlation coefficient of $+1.0$ for the two factors would mean that an estimate of accomplishment based on the one factor would be approximately 100% correct. As the amount of the correlation decreases, the accuracy of the forecast declines, but not in a direct manner. A correlation of $+1.0$ may mean 100% accuracy in the estimate based on the relationship, but a correlation of $+.50$ does not mean at all that the estimate based on it will be 50% correct. A glance at the accompanying table will demonstrate this interesting fact about the correlation coefficient. The per cents of forecasting accuracy for different

values of r given in Table 22 are obtained by applying Kelley's proposed formula for the *Coefficient of Alienation* ($k = \sqrt{1 - r^2}$) and then deducting the resulting values expressed as per cents from 100%. In cases where estimates of one variable are to be made from measurements of another related variable this table will prove to be a useful safeguard.

TABLE 22

PER CENT OF FORECASTING ACCURACY FOR SPECIFIC VALUES OF r

Coefficient of Correlation	Per Cent of Forecasting Efficiency
1.00	100
.99	86
.98	80
.95	69
.90	56
.866	50
.80	40
.75	34
.70	29
.65	24
.60	20
.50	13
.40	8
.30	5
.20	2
.10	½

A word of warning should possibly be given here in order that the student may not become over-optimistic in the interpretation of correlations. It is better to be on the safe side when making claims for the reliability or the forecasting power of a test. Much irreparable damage has been done to the cause of educational measurements by the unqualified and exaggerated statements of misinformed individuals. This damage has taken two forms. As a result of frequent misinterpretation of the measures of relationship many highly ineffective tests have been accepted and used widely. Many users of educational tests have discovered for themselves the inefficiency of the devices recommended to them and as a result of their experiences will have nothing to do with tests.

The following illustrations and practical interpretations of typical correlation coefficients representative of the sort obtained from educational data have been gleaned from a number of sources. They are offered here for whatever guidance they may give to the student or teacher in making a critical and conservative interpretation of correlation data based on test results.

109. Practical Uses of Correlation Method. The classroom teacher and the student of measurement will find the

Value of r	Educational Situation	Interpretation
+.96	Relation of two forms of a long, analytical reading test for high school students.	Evidence of unusually high reliability of measurement; treat scores with confidence.
+.90	Relation of two forms of a 45-minute group mental test.	Evidence of marked reliability.
+.80	Repetition of same form of a group test of mental ability after a lapse of one semester.	Evidence of a marked relationship; considerable prognostic power even after lapse of a long interval.
+.65	A composite of three separate estimates by same teacher of the ability of a class of 35 students to respond to an objective test in American History.	Evidence of some relationship but of limited use for prognostic purposes.
+.50	Scores on a good group intelligence test and the class grades of a class in first-year algebra.	A very slight relationship of no practical value for forecasting purposes (only 13% effective).
−.24	Chronological ages of pupils in a given grade and achievement scores on an objective test.	Negative relationship of an indifferent sort. Merely shows a very slight tendency for the younger and hence more able pupils in a grade to achieve at a higher level than the average.

reatest opportunity to use correlation techniques in connec-on with the construction and analysis of objective tests and ı the critical selection of standardized tests. The validity f a test may be expressed statistically in terms of the corre-ıtion of the instrument with some other criterion, such as nother measure whose validity is known. The determina-on of the reliability of a test, or the consistency with which measures the qualities it measures, is almost entirely a

matter of correlation method. Mastery of these uses of the correlation techniques will make the teacher a better critic of commercial standardized materials as well as a more independent and efficient builder and critic of teacher-made tests for classroom use. Such mastery can come only from extensive practice on problems calling for the use of these skills.[19] Students who are interested in the theory of correlation or in the use of correlation methods in more critical research with tests are referred to the many excellent textbooks on statistical methods now available.

Problem-Exercises

Problem 14

Prepare a correlation table of the accompanying pairs of scores representing scores on the odd and even exercises in a 100-item objective examination. Use a step of 3 on both axes. Compute the correlation coefficient of these odd-even measures (called the half-half correlation) as a basis for later work in determining the reliability of the examination. (Answer : $r_{\frac{1}{2}\frac{1}{2}} = .956$.)

Student	Score on Odds	Score on Evens	Student	Score on Odds	Score on Evens	Student	Score on Odds	Score on Evens
1	47	45	13	29	28	24	20	17
2	45	45	14	26	31	25	19	21
3	44	42	15	26	30	26	17	17
4	41	44	16	26	26	27	15	18
5	40	38	17	25	28	28	15	13
6	37	35	18	24	25	29	14	12
7	35	43	19	23	24	30	12	16
8	35	32	20	23	20	31	12	14
9	34	40	21	23	17	32	10	13
10	34	26	22	20	25	33	8	3
11	31	37	23	20	19	34	6	10
12	30	33				35	2	6

[19] See Problems 14 and 15 above and on following page, or Problems 22, 23, 24, 25, 39 and 40 in Greene's *Work-Book in Educational Measurements*, or some other source of similar problems.

Problem 15

Compute the coefficient of correlation (r) of the first and second scorings of the reading-test papers for the pupils listed here. The results will be an index to the objectivity of the test. To indicate perfect objectivity in the test the correlation should be approximately perfect (+.998). Is it?

Pupil	First Scoring	Second Scoring	Pupil	First Scoring	Second Scoring	Pupil	First Scoring	Second Scoring
1	36	36	15	31	31	29	23	23
2	35	35	16	31	30	30	23	21
3	35	35	17	31	29	31	22	22
4	34	36	18	30	32	32	22	22
5	34	35	19	30	29	33	21	21
6	34	34	20	29	30	34	20	20
7	33	34	21	29	28	35	20	19
8	33	33	22	28	29	36	19	19
9	33	33	23	27	27	37	18	18
10	33	32	24	27	26	38	17	17
11	32	33	25	26	25	39	16	17
12	32	32	26	25	25	40	15	17
13	32	32	27	24	23			
14	31	31	28	23	25			

V. Assignment of Relative and Absolute Ranks

110. **Need for Method of Ranking.** Test scores are given meaning only when they are considered in relation to some standard or norm of achievement. In working with test scores it often becomes desirable to make the achievement of other pupils in the group the basis for the comparison. The relative achievements of pupils may be compared by the simple process of assigning ranks or positions to their scores in accordance with their magnitude. That is, the pupil making the highest score is given first position in the class, the pupil making next highest is given second position, etc. This is called the assignment of *relative rank.*

The importance of resorting to some device such as the ranking of scores to take care of the fact that scores from different tests or examinations are not comparable may be easily demonstrated. Any teacher readily realizes that a mark of 80 on an examination or a test may mean almost

anything. It means one thing when there are no other marks of 80 given in the class. It means quite another thing when there are thirty marks of 80 given in the class. It may also be suggested that it means another thing depending upon how many pupils there are in the class. If the 80 is the highest grade or score made in a class of 50 pupils such a score is undoubtedly a superior mark. If the 80 is excelled by thirty pupils in the class of fifty pupils it means that the individual ranks 31st in a class of fifty. This point is discussed further in connection with the explanation of absolute ranks.

111. **Assignment of Relative Ranks.** *Relative ranks are positions assigned to pupils or scores in terms of the array of which they are a part.* The problem of ranking is merely that of assigning positions to the scores as arranged according to size. The largest score is given first position; the next largest, second; etc. The only difficulty involved appears in the case of two or more scores of the same size. The following illustration (Table 23) will make clear the method of handling this situation.

TABLE 23

RELATIVE RANKS

Pupil . . .	A	B	C	D	E	F	G	H	I	J	K	L
Score . . .	46	44	44	41	40	38	38	38	37	31	29	28
Rank . . .	1	2.5	2.5	4	5	7	7	7	9	10	11	12

Pupil A, with a score of 46, the highest score when the measures are arranged in descending order of magnitude, is assigned first rank. Pupils B and C, both with scores of 44, would normally be assigned second and third places but it is impossible to assign second place or third place to one rather than the other, so average rank, or a rank of 2.5, is assigned to both. The relative numerical position of each succeeding score is held, however. Pupil D, with a score of 41 being the fourth individual in the series, is assigned fourth place. The position assigned to the pupil with the lowest score

should coincide with the number of cases in the series except in case the last scores are tied.

Care should be taken in the ranking of individuals from scores in order to avoid errors in the assignment of positions. The practice of recording such scores on small cards which can be sorted and arranged in order will eliminate many mistakes if the series runs into many cases.

It must be clear to the student that the assignment of relative rank or position to pupils earning certain test scores actually covers up something of the true situation. For example, the differences in the magnitudes of scores are submerged by the assignment of ranks. To illustrate, in the case of the data used in Table 23, a difference of 6 points on the test scale (from 31 to 37) makes a difference of only one rank position (from 9th position to 10th). A difference of only one test point (from 28 to 29) also makes a difference of one rank position. This is a point which should be kept in mind when using the method of relative ranks. Thus ranking tells us that one pupil is ahead or behind another pupil but fails to tell how much in terms of actual test scores he is ahead or behind the other pupil.

Problem-Exercise

Problem 16

Assignment of Relative Ranks

Rank and assign relative positions to the scores in Problems 1, 2, and 3.

112. **Assignment of Absolute Ranks.** The usefulness of relative rank is somewhat limited by the fact that it takes no account of the actual level at which the accomplishment takes place. A person ranking 32 in a group of 35 has a very low relative rank in this group. However, if he ranked 32 in a group of 250 the significance of his accomplishment would be greatly changed. *Percentile ranks* take this point into account by reducing the ranking to a basis of 100 units. A percentile score of 75 means that for the measures under consideration the individual assigned that absolute rank made a score which exceeds 75% of the individuals of his group without regard to the number of cases it contains.

The student will recognize this 75th percentile as a measure with which he has already had some experience. The 75th percentile of a distribution is exactly the same thing as the third or upper quartile (Q_3). The same methods by which the 25th, 50th, or 75th percentiles are obtained may be applied to the computation of any designated percentile. It is often useful to determine the 10th, 20th, 30th, 40th percentiles. These are called *deciles*, and are computed in exactly the same manner as the median and the quartiles.

TABLE 24

COMPUTING DECILES

Class Intervals	f	Cumulative Frequency[20]		Percentile Scores
70.5–73.5	1	25	90% of 25 = 22.5	90th = 63.75
67.5–70.5	0	24	80% of 25 = 20.0	80th = 60.50
64.5–67.5	1	24	70% of 25 = 17.5	70th = 58.00
61.5–64.5	2	23	60% of 25 = 15.0	60th = 55.50
58.5–61.5	3	21	50% of 25 = 12.5	50th = 54.25
55.5–58.5	3	18	40% of 25 = 10.0	40th = 53.00
52.5–55.5	6	15	30% of 25 = 7.5	30th = 50.25
49.5–52.5	2	9	20% of 25 = 5.0	20th = 46.50
46.5–49.5	2	7	10% of 25 = 2.5	10th = 39.00
43.5–46.5	2	5		
40.5–43.5	0	3		
37.5–40.5	1	3		
34.5–37.5	1	2		
31.5–34.5	0	1		
28.5–31.5	1	1		
	$N = 25$			

113. Computing the Percentiles (Deciles) in a Frequency Distribution. The calculation of the 10th, 20th, 30th, etc. percentiles involves exactly the same general procedure as that followed in computing the median. The only significant difference is, of course, in the number of cases used in

[20] The cumulative frequencies which are obtained by taking the cumulative total of the cases from the bottom of the distribution toward the top are given in this table for convenience in computing the necessary interpolations.

counting into the frequency distribution. In computing the median (50th percentile) exactly one-half of the scores are counted off. In computing the 25th percentile (Q_1) one-fourth of the cases are considered. In computing the 10th percentile quite naturally only one-tenth of the cases are counted off. • The interpolation into the proper step is exactly identical to the method followed in computing the median. Table 24 shows the general methods and results without detailed comment. Since the student has already done work of a similar nature no difficulty should be found in understanding the correctness of every figure in the table.

TABLE 25

INTERPRETATION OF PERCENTILES

Percentiles (Deciles)	Interpretation	Test Scores
100	Score equalled or excelled by no student..	71.00
90	Score equalled or excelled by 10% of the students..........................	63.75
80	Score equalled or excelled by 20% of the students..........................	60.50
70	Score equalled or excelled by 30% of the students..........................	58.00
60	Score equalled or excelled by 40% of the students..........................	55.50
50	Median—score equalled or excelled by 50% of the students................	54.25
40	Score equalled or excelled by 60% of the students..........................	53.00
30	Score equalled or excelled by 70% of the students..........................	50.25
20	Score equalled or excelled by 80% of the students..........................	46.50
10	Score equalled or excelled by 90% of the students..........................	39.00
0	Score equalled or excelled by practically 100% of the cases..................	30.00

The computations in Table 24 give the percentile scores reported in Table 25. The score assigned as the 100th percentile is the highest score obtained by any pupil on the test.

used in the table given as an illustration. In this case it is assumed to be 71. The percentile score assigned as zero (0) is the lowest score made by any pupil on the test. In this illustration it is assumed as 30.

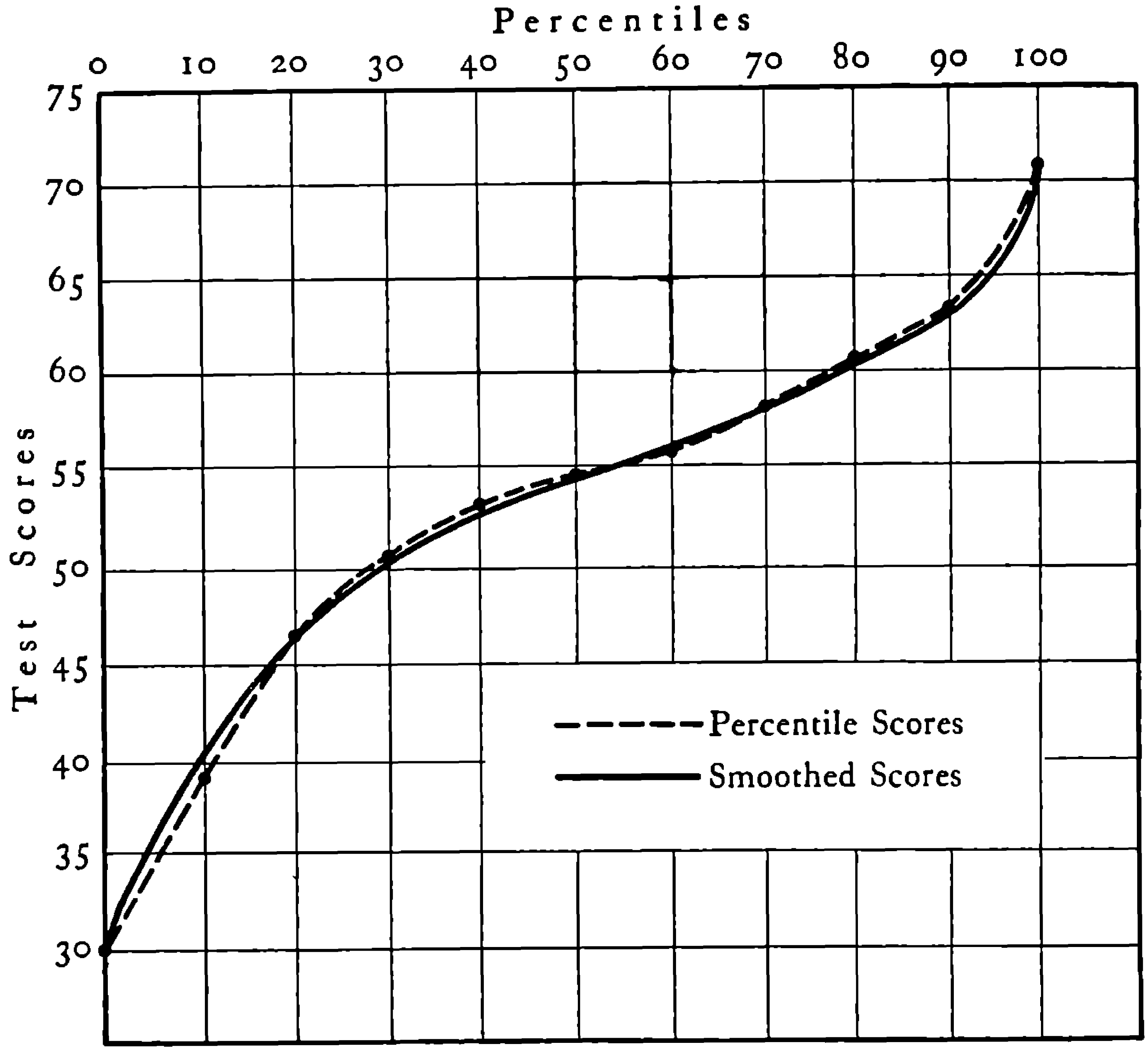


FIGURE 9. PERCENTILE GRAPH

The interpretation of percentile scores frequently gives some trouble to the worker inexperienced in their use. Figure 9 is a graphic presentation of the percentile scores given in Table 25. This figure shows the characteristic curve (ogive) resulting from the use of percentile scores. The heavy solid line in the figure represents the results of an arbitrary smoothing of these percentile scores. This smoothing process is frequently used when percentile scores are based on fairly large populations and are set up as tentative norms for the interpretation of the tests. Further reference to norms of this type will be made in Chapter IX of this book.

Problem-Exercise

Problem 17

From the frequency tables prepared in connection with Problems 1 and 2, compute the scores corresponding to the following percentiles in each distribution: 1, 10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90, and 99.

Problem 18

Follow the procedure outlined in Figure 9 and prepare a chart showing the percentile (ogive) curve for the values calculated in Problem 17.

VI. Summary

This chapter presents a non-technical discussion of a few of the common statistical tools which the beginning student in educational measurements should find useful in the analysis and interpretation of educational test results. Discussions of four of the six major statistical techniques outlined in the introductory paragraph of this chapter are presented. The fundamental principles of the grouping and tabulating of test scores are stated and illustrated. The need for measures of dispersion is shown. The quartile deviation and the standard deviation are explained in some detail and practical applications of these measures to problems of test analysis are made. The general meaning and the methods of correlation are given, along with a few definite hints concerning the interpretation of correlation coefficients. The practical uses and meanings of the ranking of test scores on both the relative and the absolute basis are discussed. The two remaining problems, dealing with the derivation and interpretation of test norms and standards, and the use of simple graphic methods of presenting the results of statistical analysis are reserved for treatment in the following chapter.

There has been no attempt to make this chapter a complete discussion of all the interesting or even useful statistical techniques. To do this would require a volume in itself. As a matter of fact the brevity of the treatment makes it impossible to present a sufficient number of illustrations and problems to give the inexperienced worker adequate experience with these statistical problems. Real mastery of these

technical skills can come only through their repeated and continuous use. Technical facility will not come from a reading of this chapter, but an appreciation of the problems involved and the meaning of the novel notations and vocabulary should come from repeated study of the discussions and illustrations. Without the type of quantitative thinking emphasized in this chapter the student of education is forced to use vague methods in his appraisals of educational data.

Selected References

Garrett, H. E., *Statistics in Psychology and Education.* New York : Longmans, Green and Company, 1926.

Good, Warren R., *The Elements of Statistics.* Ann Arbor : The Ann Arbor Press, 1933.

Greene, H. A., *Work-Book in Educational Measurements.* New York : Longmans, Green and Company, 1930.

Holzinger, Karl J., *Statistical Methods for Students in Education.* Boston : Ginn and Company, 1928.

Holzinger, Karl J., *Statistical Tables for Students in Education and Psychology.* Chicago : University of Chicago Press, 1925.

Holzinger, Karl J., and Mitchell, B. C., *Exercise Manual in Statistics.* Boston : Ginn and Company, 1929.

Kelley, T. L., *Statistical Method.* New York : Macmillan Company, 1923.

Kramer, Edna E., *Educational Statistics.* New York : John Wiley and Sons, Inc., 1935.

Lindquist, E. F., and Stoddard, George D., *Study Manual in Elementary Statistics.* New York : Longmans, Green and Company, 1929.

Odell, C. W., *Educational Statistics.* New York : The Century Company, 1925.

CHAPTER IX

INTERPRETING THE RESULTS OF TESTING

The purpose of this chapter is the consideration of the following important points in the interpretation of test results:

a. Meaning of a test score.
b. Giving meaning to test scores through the use of norms and standards.
c. Differences between norms and standards.
d. Functions of grade norms in the interpretation of test scores.
e. Age norms in the interpretation of test scores.
f. Meaning and function of "derived scores."
g. Interpreting test scores through the use of quotients.
h. Comparing scores from different tests through the use of derived scores.

I. Test Scores

The problems of summarizing test scores and of interpreting the results revealed by these summaries are very closely related and, were it not for the length and detail of the discussion which they require, should probably be considered in a single chapter. The preceding chapter, Chapter VIII, deals with four of the six major problems of statistical procedure as related to test analysis. The content of this chapter concerns itself almost entirely with the last two of these problems of test interpretation.

114. **Meaning of a Test Score.** Test scores are valuable to the classroom teacher to the extent that they can be interpreted. It is therefore important to define clearly what is meant by a test score. In order to accomplish this, two or three new concepts require explanation. In the first place, *a test score is a numerical expression of performance* on the part of an individual. Sometimes the test score is merely

the number of exercises responded to correctly. Again it may be an arbitrarily defined scale value. But whatever its form, its function is to reveal in a quantitative way the performance of an individual as he responds to stimuli given under certain conditions. This leads to the second concept involved in the meaning of a score. The test score is an evidence of performance. Performance, the response of the individual to the test situation, is taken to mean in educational measurements the expression of ability operating under certain conditions. *Performance may be thought of as ability + conditions.* Scores on tests are definitely influenced by conditions. The pupil may make a poor score because he does not have the ability to do better — may not know the facts, etc. On the other hand he may make a low score because of certain physical conditions, illness, discomfort, hearing, sight, illumination, a broken pencil, an indifference for the subject, a dislike for the teacher or examiner, a failure to give attention to and to comprehend the directions, etc. Any one of these or a dozen other factors may affect the score. Accordingly, there is the possibility and even likelihood of a serious error in the assumption that a test score is a direct evidence of ability. The conditions under which the performance takes place must be known before it is safe to infer ability from performance.

Ability, as an abstract concept, may be defined as the power to do. Power to do, to respond to stimuli and to situations, is the product of training and experience. *Ability may be thought of as Capacity + Training.* This suggests that unless training and native capacity factors are known, inferences as to abilities may be misleading. This point becomes particularly serious in the interpretation of mental test results, for it is a common practice for users of mental tests to infer innate capacity (mental ability) from performance scores. The real seriousness of this type of uncritical inference may be seen by comparing the interpretations of an achievement test score and a mental test score. Both are basically expressions of performance. Equal abilities may be inferred from equal scores in both types of tests if and when the conditions under which they are given are all definitely under control. While it is difficult to make sure that

all physical and physiological factors are adequately controlled in a testing situation, it is possible to regulate most of the mechanical conditions within reasonable limits. The significant point to note here, however, is the fact that users of achievement tests stop with an inference of equal ability from equal performance scores, but users of mental tests are obliged to make a further inference. In the interpretation of mental test results it is common practice to infer equal native capacity from apparent evidences of equal abilities. The fallacies in this argument and the dangers of this step must be readily apparent. *Equal capacities may be inferred from performance scores only when there is direct and positive evidence of two things; first, that the conditions under which the testing took place were identical and equally well controlled; second, that the training opportunities of the individuals compared have been equal.* The mechanics of testing now make it fairly easy to control testing conditions. The second factor represents a real stumbling block in the way of an accurate and sane interpretation of the mental test results. The naïve manner in which some makers and many users of mental tests assume equality of learning opportunity, and hence equal capacity from equal performance scores on mental tests is one of the things which has made many teachers and students skeptical of their value.

It is possible that the foregoing discussion of the meaning of a test score may appear to indicate that it is impossible to give meaning to any kind of test score. Such is not the intention, even though the purpose here is to emphasize the need for a conservative attitude in test score interpretation. In the long run the more that is known about the variables underlying test scores, the more critical must the user become. The greatest damage that has been done to the field of educational measurements in the past has come as a direct result of carelessness and ignorance on the part of users of tests, and their tendency to draw unwarranted conclusions from the results. The teacher should be able critically to select suitable tests and scales for classroom use, control the mechanical conditions of their administration, and draw sane and defensible conclusions and inferences from the results.

115. **Giving Meaning to Informal Test Scores.** The user of educational tests in the classroom is confronted with two types of test data for interpretation. The first type, and undoubtedly the more common of the two, deals with the results of informal, teacher-made tests. The results from these home-made tests are in turn of two types, the subjective scores assigned by teachers to pupils' responses to essay-type tests, and the performance scores resulting from informal objective examinations. While something can be done to improve the interpretation of the relatively unreliable marks assigned to the discussion-type exercise, much more is possible in the complete and accurate interpretation of the performance scores which are the result of long and reliable objective examinations. Since one of the major functions of the standardization of a test is the establishment of meaning for the test scores, many additional types of interpretation are thus made possible for the standardized instrument.

II. Test Norms and Standards

116. **Meaning of Standardization.** Early in the history of the development of objective testing practically all that the development of a standardized test required was to give a few test exercises to a hundred or more pupils in different school systems. The results were then compiled and submitted as norms. The standardized test differed from a reasonably good informal objective test only in the fact that the former had been tried out with more pupils in a larger number of different classes. Test standardization as it is now interpreted, however, means much more than the mere derivation of norms, although the existence of norms is still one of its chief distinctive features.

In terms of present-day test-construction practices the standardization of a test involves a long period of experimentation with a large body of subject-matter exercises. After the subject-matter field to be tested has been decided upon, there is the very difficult problem of selecting the more important areas of this field to be sampled. Many times experimental evidence must be secured before it is possible to decide upon the best type of test exercise to use. Even

then many of the exercises prepared in preliminary form for the test are found to be badly stated or to be totally unsuited for the type of test to be constructed. Usually from four to six times as many exercises must be prepared in the preliminary work with a test as will appear in the test itself when in final and standardized form. Special care must be taken to see that a suitable range of difficulty is provided in the items, and that multiple items covering certain of the more important skills are prepared in parallel form so that these items may be adequately sampled in the several forms of the test which must be prepared. After the exercises themselves have been written in preliminary form they must be tried out under experimental conditions in typical classrooms for the purpose of discovering the faulty or ambiguous items and the relative difficulty of the several items. From the results of this preliminary use of the exercises two or more roughly scaled forms of the tests may be set up for further experimental use. From the results of this second trial it is usually possible to equate the two sets of test items quite closely by shifting hard and easy items from one form to another until approximate equality is reached. Then the tests are ready for a further trial in a large number of representative classes for the purpose of further equating the forms and establishing norms. It is thus apparent that while standardization is only one of the final steps in the preparation of a carefully made test, it is this extensive sampling of the results of the use of the test in many classrooms which affords the basis for the assignment of meaning to the test scores.

117. **Test Norms.** Many of our present-day standardized tests began their existence as informal objective examinations. In fact, many informal examinations of the objective type meet all other criteria of standard tests except that they do not have norms for the evaluation of their scores. Standardized tests are characterized by the fact that they are commonly accompanied by norms or tables of typical scores representative of the accomplishment which may be expected from classes similar to those used in the standardization program. Norms are tables of information necessary for the interpretation of the test scores, and are obtained by giving the particular test to a large and representative sampling of

pupils in the same grades and of a type similar to the groups in which teachers will use the tests. To the extent that the sampling is distributed over a large population in typical school situations, and the conditions under which the tests are to be administered are rigidly followed, the norms furnish a reliable and useful basis for interpretation.

118. **Norms and Standards.** The use of the term *standardized* in the discussion of tests of the type for which norms are provided has led to the development of a careless tendency to use the words "standards" and "norms" as synonyms. The process of securing the data for the critical analysis of tests and the derivation of suitable norms is properly known as *standardizing.* However, *the term standards, when used to refer to levels of pupil-achievement, implies an ultimate goal to be achieved.* Standards may not actually be reached by any individual, but they are levels of achievement toward which to strive. *Norms are the levels of achievement which typical pupils actually attain.* When considered in the light of these definitions, it is clear that there are few tests which are accompanied by standards. It might be more nearly the truth to call the process of securing these comparative scores interchangeably known as norms and standards by the more descriptive name of "normalizing."

Possibly one of the best illustrations of the differences between standards and norms is found in the field of arithmetical computations. The standard of arithmetical accuracy is naturally 100% for most such computations containing, error are useless. However, the actual norm of arithmetical accuracy of computation on a well-known test is from 65% to 70% for the junior high school grades. That is to say, pupils of these grades work these particular kinds of arithmetical examples with an accuracy of from 65 to 70 per cent instead of the desired ultimate goal of 100% accuracy.

119. **Use of Test Norms and Standards.** Norms are distinctive features of standard tests. The distinctions between norms and standards are apparent. The prevalence of norms rather than standards for most educational tests forces us in the interests of consistency to use the term "norms" most

frequently in this discussion. It should be remembered that many times a norm (an average or typical achievement) is not a respectable total of achievement. Ambitious teachers will recognize this limitation of test norms as usually interpreted and encourage their pupils to strive for a superior type of accomplishment.

A number of specific uses of test norms and standards are listed below:

1. It is by means of the norms and standards that test scores are given meaning. This is illustrated concretely by the data in Tables 26 and 30 on pages 218 and 227.
2. Norms afford the basis for the evaluation of instruction. If the achievement of given school groups is consistently below the norm an investigation into the underlying causes should be made. See Figure 11, on page 224.
3. Norms provide the basis for goals of achievement which are valuable for the stimulation of pupil effort.
4. Norms provide the basis for the analysis of individual pupil's test scores for the purpose of revealing individual difficulties and weaknesses. See particularly Tables 31, 32 and 33 on pages 229 and 230.
5. Norms afford an objective basis for determining whether a pupil is working up to the level of ability which may be expected of him. See Table 39, page 242.
6. Reliable norms furnish the necessary information for the more accurate and objective classification and gradation of the pupil. See Table 35, page 233.

120. Kinds of Norms. The form in which the norm for a test is provided depends to a large degree upon the level in the school system at which the test is used. The norm is also conditioned somewhat by the nature of the test itself. Tests which are designed for use in the elementary school grades are usually accompanied by two types of norms, *grade norms*, and *age norms*. Tests intended for use in the secondary school grades are commonly provided with percentile and grade norms only. Age norms do not seem to be particularly useful in the high school and college levels, due to the fact that so many factors other than age operate to affect achievement. Then too the curve of mental growth flattens out very rapidly after the fifteenth or sixteenth year so that the increments of growth in achievement from age

to age at the upper levels are relatively not significant. In place of the age norms for secondary school and college tests the common practice today is to provide detailed tables of percentile equivalents for the point scores. These tables permit a very satisfactory interpretation of the test scores at these grade and age levels.

In the lower grades, age as well as grade norms usually accompany achievement tests. In general, the type of norm is determined by the method of grouping the scores when the tabulations for the norms is made. If the pupils are grouped by grades, without respect to age or school progress, the resulting norms are grade norms. If the pupils are classified in accordance with a chronological or mental age-scale as the basis for the tabulation, the resulting norms are age norms.

In the derivation of grade norms for standard tests it is a common but not universal practice to express the norms in terms of end-of-the-year achievement. In any event, it is desirable to have the norms clearly indicate the period they are designed to cover. In the derivation of the grade-equivalents from the grade norms, the norms are adjusted in such a way that the approximate progress of the pupil through the grade is indicated by the grade-equivalent assigned to his score. For example, the seventh year grade end-of-the-year norms for a certain test might be 120 points, and the eighth grade end-of-the-year norm 140 points. A score of 130 points would be treated as equivalent to achievement half-way through the eighth grade or 8^5. A score of 140 would be recorded as 8^{10}. The exponent in each case represents the number of tenths of a grade of achievement revealed by the score in terms of the grade norms.

The problem of establishing age norms for tests has caused a great deal of difficulty. An early practice in the preparation of age norms involved the re-grouping of all pupils used in the grade tabulation into chronological age groups regardless of grades. The test scores of these chronological age groups were then tabulated, and the means or medians computed. These results were then used as the basis for setting up tables of the point scores corresponding to the several age groups. It soon became apparent, however, that many

factors other than age were operating to influence the average achievement of pupils grouped in grades. Such factors as over-ageness, retardation, a lack of balance between retardation and acceleration were all present. It was found, for example, that while the average chronological age of a seventh grade pupil might be 13 years and 6 months at the end of the school year, the average test score of pupils 13 years and 6 months of age was not at all the same as the end-of-year score for the seventh grade. This caused much confusion in the interpretation of grade and age data. Obviously some method of taking care of this problem was required.

Some very useful evidence on the influence of such variables as irregular school progress, accelerated or retarded mental growth, and over-ageness and under-ageness has been presented recently in a study by Dr. J. R. Crawford.[1] This investigation indicates that test norms based on unselected cases are definitely affected by the relative balance of under-age and over-age pupils. The actual achievement of the under-age pupils is significantly superior to the over-age groups in a given grade, as might be expected. For the makers of standardized tests the useful implication from this study lies in the fact that it makes very apparent the real need for differential norms, or norms which will take into account wide differences in mental ability or school progress within the grade.

One interesting development in the preparation of test norms which has come as a by-product of the Crawford study is the use of the mental ages rather than the chronological ages of individual pupils in the derivation of age norms. The results of one application of this procedure in the derivation of norms for a new test[2] are shown in Table 26. This table is somewhat novel in that it reveals the types of reading scores which may be expected from pupils of different mental age levels within the grades. It sets forth with unusual clearness the fact which apparently has not always been understood by teachers that a pupil of *twelve years mental age* in the *fourth grade* is not at all the same individual so

[1] Crawford, J. R., *Age and Progress Factors in Test Norms.* Iowa City : University of Iowa Studies in Education, No. 4, Vol. IX, June 15, 1934.

[2] *Iowa Silent Reading Tests : Elementary Test.* Yonkers, New York : World Book Company, 1933.

Table 26

DIFFERENTIAL MENTAL-AGE-GRADE NORMS
IOWA SILENT READING TESTS[3]

Reading Age		Grades					
Yr.	Mo.	4	5	6	7	8	9
17	0						187
16	9						182
16	6					173	178
16	3					168	174
16	0					164	170
15	9					159	166
15	6				152	155	161
15	3				148	150	158
15	0				144	146	153
14	9				140	141	149
14	6			133	136	138	144
14	3			129	132	135	140
14	0			124	128	131	136
13	9			119	124	127	132
13	6			115	120	123	128
13	3			110	116	118	124
13	0		90	106	111	114	119
12	9		88	102	107	110	115
12	6		86	98	102	106	110
12	3		83	93	98	102	
12	0	75	81	89	92	98	
11	9	72	79	85	89		
11	6	69	77	82	86		
11	3	66	74	77	80		
11	0	63	70	73	76		
10	9	60	66	70			
10	6	57	63	67			
10	3	53	60	63			
10	0	48	56	60			
9	9	43	52				
9	6	37	47				
9	3	33	42				
9	0	29	37				
8	9	27					
8	6	25					
8	3	24					
8	0	23					

[3] Reproduced with special permission of the publishers, World Book Company, 1933.

ar as achievement is concerned as the *twelve year old in the* *ifth* or *sixth grade.* A reference to the data in Table 26 will nake this quite apparent.

This table makes it possible to interpret these reading test cores in terms of the grade location and mental age of the ndividual pupil, or by reversing the reading it permits the nore exact assignment of reading ages in relation to a pupil's rade location. The scores printed in black face type in the able represent the normal achievement of the normal mental ge group for the grade.

TABLE 27

SAMPLING OF SCORES ON AN ENGLISH GRAMMAR INFORMATION TEST, EXPERIMENTAL EDITION

Mid-point of Class-Intervals	Grades			
	9	10	11	12
75			1	1
72			2	4
69		1	5	6
66	1	2	6	4
63	3	2	7	8
60	3	5	10	12
57	7	8	11	13
54	10	12	10	19
51	14	7	14	14
48	13	14	19	23
45	15	12	27	17
42	18	25	26	16
39	23	20	22	17
36	27	17	17	10
33	25	18	10	14
30	28	19	12	12
27	24	18	10	10
24	21	15	12	8
21	16	10	9	2
18	17	8	3	4
15	12	11	1	2
12	11	8	1	
9	5	2		
Total	293	234	235	216

121. **Derivation of Norms.** The grade norms established for most of the commonly used achievement tests are based on the median scores obtained by giving the tests to large groups of pupils in each grade. Such norms provide the basis for the interpretation of class scores as well as of individual accomplishment. They are commonly considered as defining the degree of ability which a pupil or a group of pupils should possess. In some instances the idea of perfection is associated with a norm. However, present methods of standardizing a test should make clear that we are not justified in thinking of norms as statements of what ought to be. As they are now derived, norms should be interpreted as representing the average achievement of certain groups of pupils under present school conditions.

TABLE 28[4]

PERCENTILE NORMS, SMOOTHED

(Based on Data in Table 27)

Percentile Scores	Grades			
	9	10	11	12
99	63	66	70	73
90	51	55	60	63
80	45	48	53	58
75	42	46	50	56
70	40	43	48	54
60	36	40	45	49
50	33	36	43	46
40	30	32	40	42
30	26	28	36	39
25	24	26	33	36
20	22	24	30	33
10	16	19	24	27
1	9	12	15	18

The derivation of grade norms is accomplished by the simple process of giving the test to be standardized to a large and unselected group of pupils in the grades for which

[4] A further illustration of the steps in deriving grade norms for test material is given in Problem 29 of *Work-Book in Educational Measurements*, by H. A. Greene.

the norms are desired under the conditions which are to operate in the later use of the tests. The actual steps in the process are illustrated in Table 27. This table represents the frequency distributions of the scores made by the pupils in grades 9, 10, 11, and 12 for which the norms are desired. The data are a few of the actual cases used in deriving the norms for a new information test in English grammar.[5] In this case the norms are arranged to show not only median achievement per grade, but also each decile of achievement (i.e., each tenth percentile, as 10th, 20th, 30th percentile, etc.). The norms themselves based on these distributions (Table 27) are shown in Table 28, on page 220. Figure 10, page 222, shows in graphic form the median, 25th percentile, and 75th percentile of achievement for this test.

The steps in the process of deriving grade norms for a test represent in a general way the procedure involved in securing age norms, or progress norms. Percentile norms are not commonly reported for the different age groups. They afford very helpful supplementary information for the interpretation of grade norms, however.

122. **Meaning of Test Norms.** In the use of standard tests for general survey purposes or for purposes of group comparisons grade norms ordinarily afford the chief basis for interpretation. The meaning which may be attached to such interpretations, however, depends upon a number of factors which the norms themselves do not take into account directly. For example, it is assumed that when tests are being given for the purpose of deriving norms of achievement the exact conditions operate which are to be maintained when the tests are given in the classroom. The subjects used are grouped according to the training they have had in the field of the test, or according to the experience they have had in the taking of tests. Certain time limits of work are established. Certain statistical procedures are utilized in analyzing the data. If comparisons with norms affected by these conditions are to be valid the results compared must be secured under identical conditions. The specific directions accompanying standard tests must be followed exactly if compari-

[5] Cram, Fred D., and Greene, H. A., *Iowa Grammar Information Test*. Bureau of Educational Research and Service, University of Iowa, Iowa City, 1935.

sons with norms are to be made. The groups of subjects (pupils) must themselves be comparable in terms of training, experience, and criteria of classification. Statistical techniques similar to those used in setting up the norms must be employed in making the comparison. That is, if the

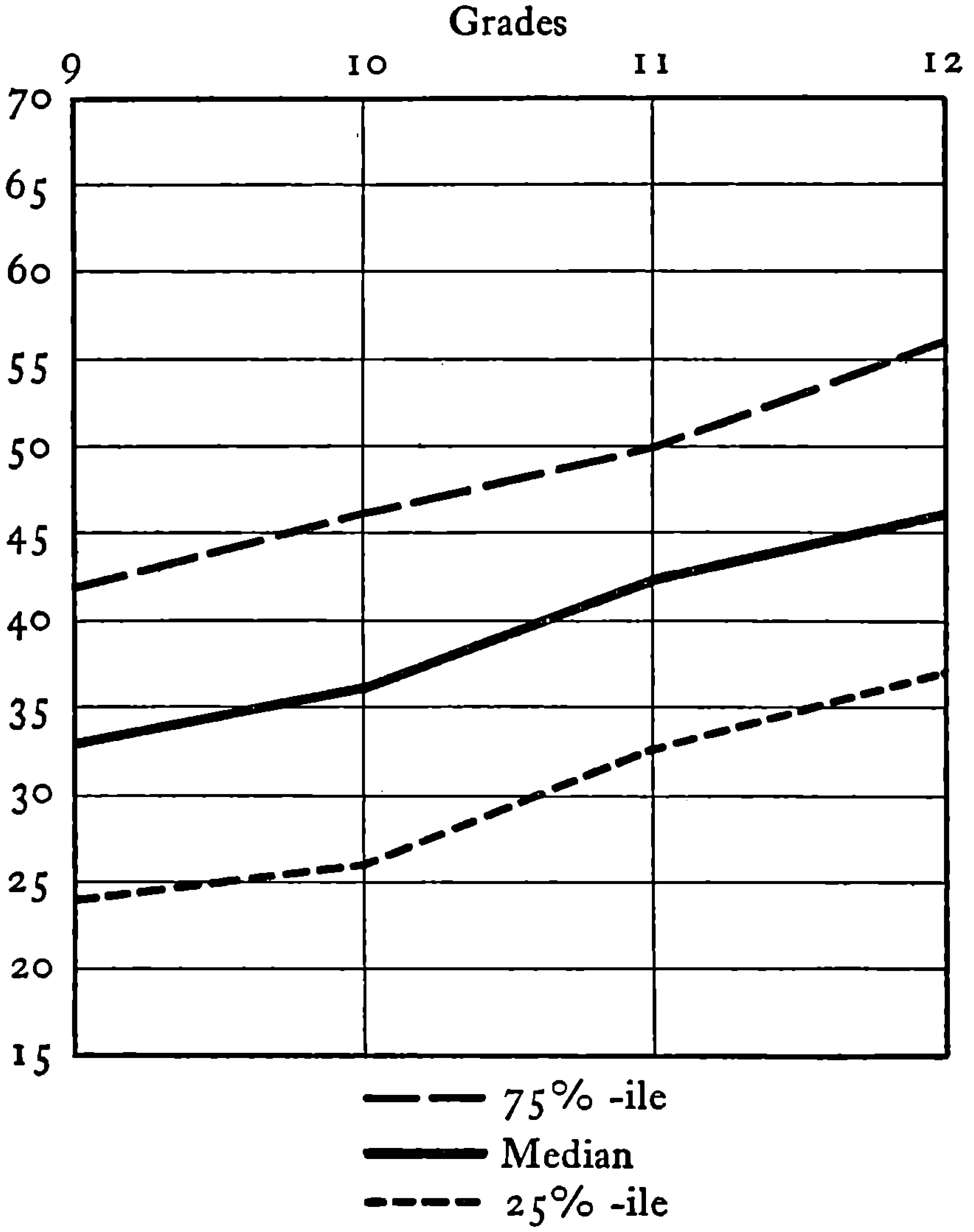


FIGURE 10. GRAPHIC PRESENTATION OF THE MEDIAN, 75TH AND 25TH PERCENTILES GIVEN IN TABLE 28

median achievement of the standardization group is used as a norm then the median scores of the classes to be compared with the norms should be used. If a percentile score such as the 60th percentile is used as the standard, then a similar percentile should be used for comparison. If the norms are established for a particular period or amount of school train-

ing, account of this must also be taken in making any interpretations of results.

Since norms reflect average or typical achievement it is, of course, obvious to the student that the school system which is above the average in such matters as the social and mental level of the population from which it draws its pupils, the quality of the physical equipment, the training and qualifications of its teachers, etc., should not be satisfied with merely normal scores. Norms are for the guidance of the average. Those who wish to excel in their achievement should establish higher goals for themselves.

123. **Grade Norms.** The functions of grade norms for group comparisons may be illustrated by data from the *Ayres Measuring Scale for Handwriting.*[6] The solid line in Figure 11 represents the handwriting quality which may be expected from typical children in grades three to eight inclusive. The broken line in the graph shows the grade medians of writing quality for these grades at the middle of the year as revealed in a state-wide survey of handwriting involving 38,890 specimens collected in Iowa in 1927.[7] The comparisons show that the Iowa state-wide results in handwriting quality as revealed by this survey are quite significantly and uniformly below the standards of quality which accompany the scale. This graphic presentation points out the difference between norms and specific achievement, or between standards and norms, depending upon the point of view. If in this figure the solid line is treated as representing standards, the broken line, being based on a sampling of several thousand cases on a state-wide basis may be considered as representing norms of handwriting quality for the state.

124. **Age Norms.** Grade norms expressed as medians and percentiles afford very useful comparisons in the high school grades where the influences of the age factors are not so evident. They are, however, generally inadequate in the elementary grades. The almost complete absence of age norms

[6] *Ayres Measuring Scale for Handwriting.* Published by Russell Sage Foundation, New York.

[7] Greene, H. A., "Report of Iowa State-Wide Survey of Handwriting." Unpublished Study distributed by Bureau of Educational Research and Service, University of Iowa, 1927.

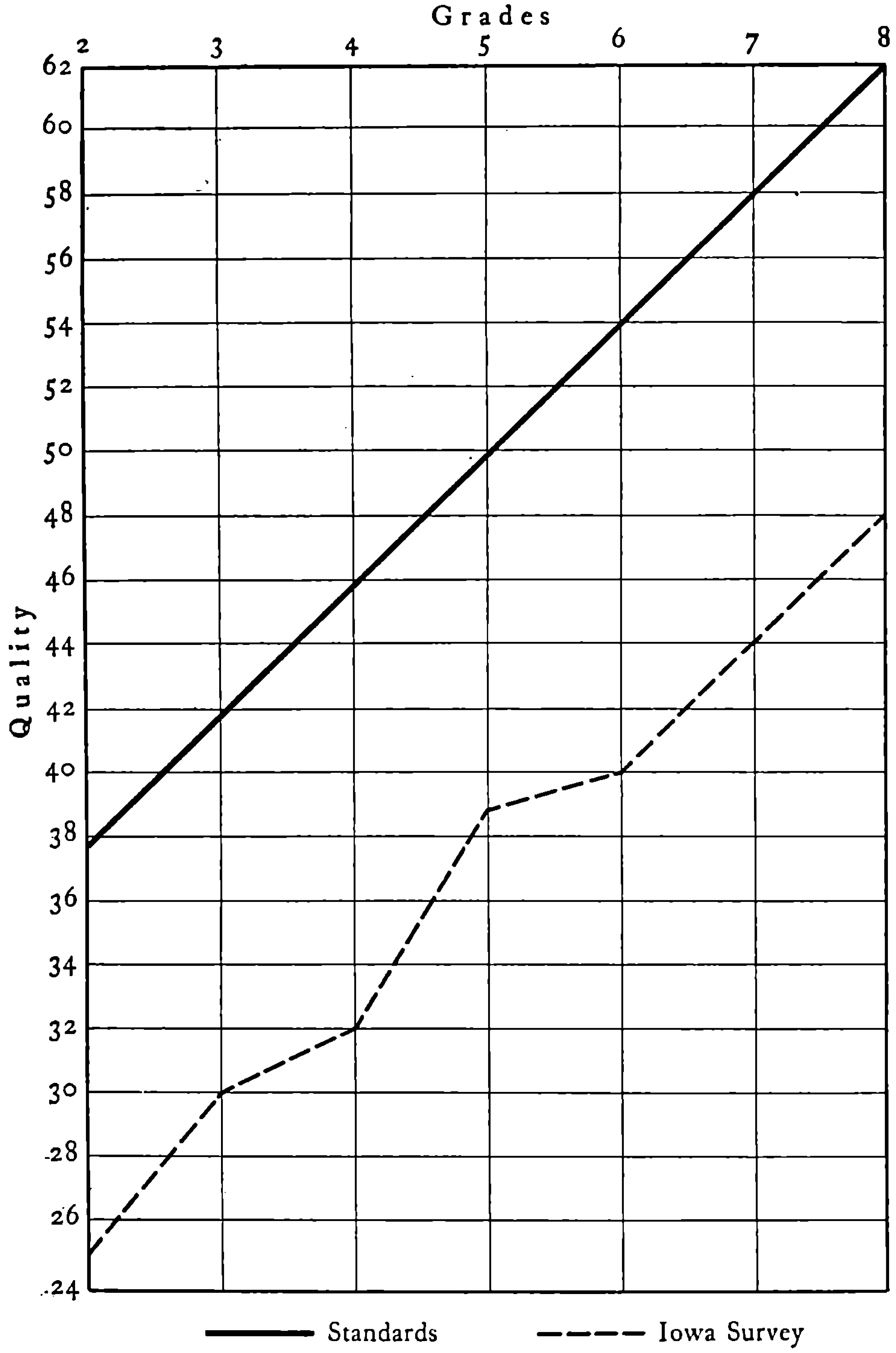


FIGURE 11. COMPARISON OF HANDWRITING QUALITY WITH GRADE NORMS (AYRES)

for tests for high school use makes it necessary to draw all illustrations in this section from elementary school data.

The inadequacy of grade norms as the basis for comparison with the test scores of pupils is nicely illustrated by the

data in Table 29. This table is based upon data similar to that presented in Table 26 except that it shows a distribution of the sixth grade pupils only on a chronological age basis within the grade. The median comprehension scores on the *Iowa Silent Reading Test, Elementary,* are shown for each of the six-month age groups found in the sixth grade. The medians reveal a fact which is well-known to every experienced classroom teacher, that the older pupils within a given grade make an average or typical score which is lower than that

TABLE 29

DISTRIBUTION OF IOWA SILENT READING COMPREHENSION SCORES BY CHRONOLOGICAL AGES

(868 Sixth Grade Pupils)

Chronological Age		Number Cases	Mean Total Comprehension Score
Yrs.	Mos.		
14	0	59	74
13	6	66	84
13	0	109	92
12	6	125	98
12	0	206	103
11	6	177	108
11	0	95	114
10	6	17	121
10	0	14	124

made by the younger pupils within the same grade. This same situation is shown graphically in Figure 12. The solid line represents the median scores made by each half-year chronological age group within the grade. In general these medians decline as the older age groups are reached. The solid horizontal line represents the median score earned by all sixth grade cases. This figure shows very definitely how the median score for this sixth grade class would work an obvious injustice on the younger and more able members of the class. A recognition of the evident inability of grade norms to cope with this situation has prompted authors of

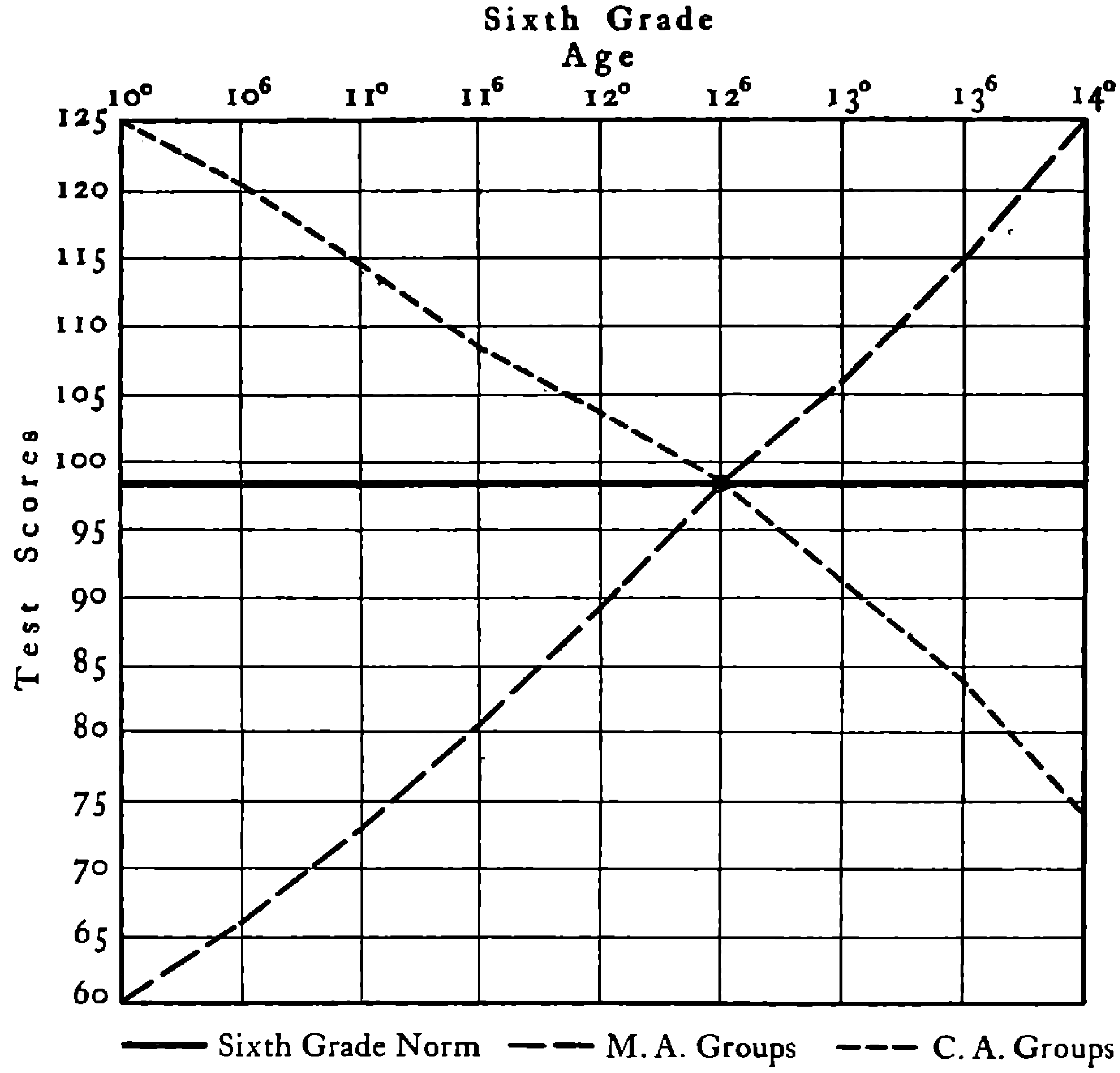


FIGURE 12. ILLUSTRATING NEED FOR REFINEMENT IN AGE NORMS

most of the modern tests to provide norms for age groups as well as grade groups.

The data presented in Table 26 illustrate a procedure recently proposed to give still more exact meaning to age norms. In this tabulation the age classification is made on a mental basis. The effect of the control afforded by this classification is to make it very obvious that an individual of a given mental age may be expected to achieve at a higher level as he is found in more advanced grades. For example, the mental twelve-year old in the fourth grade makes a typical score of 75 points (See Table 26) while fifth grade pupils of the same mental age make a typical score of 81 points, sixth grade pupils of the same mental age make scores of 89, etc. Figure 12 is an attempt to show the relative effects of classification of pupils in a given grade on a chronological age and on a mental age basis. The solid horizontal line

epresents the sixth grade norm based on all pupils used n the standardization. The broken line running from the ower left-hand corner of the graph to the upper right-hand orner indicates the scores made by the different age groups n the sixth grade classified by mental ages. The other ine running in general from the upper left- to the lower ight-hand corners of the graph shows what may be expected vhen the 868 cases are tabulated by chronological age groups. The data given in Tables 26 and 29, and in this graph (Figure 12) are considered to be rather convincing evidence of he need for more exactness in the establishment of age norms.

TABLE 30

AGE-EQUIVALENTS OF COMPOSITE SCORES OF NEW STANFORD ACHIEVEMENT TESTS, ADVANCED EXAMINATION

Composite Scores	Age Equivalents Yr. Mo.	Composite Scores	Age Equivalents Yr. Mo.	Composite Scores	Age Equivalents Yr. Mo.	Composite Scores	Age Equivalents Yr. Mo.
120	19– 2	94	14– 8	69	11– 5	44	9– 6
119	18–11	93	14– 6	68	11– 4	43	9– 5
118	18– 8	92	14– 4	67	11– 3	42	9– 4
117	18– 5	91	14– 1	66	11– 2	41	9– 3
116	18– 2	90	13–11	65	11– 1	40	9– 3
115	17–11	89	13– 9	64	11– 0	39	9– 2
114	17– 8	88	13– 7	63	10–11	38	9– 1
113	17– 6	87	13– 5	62	10–10	37	9– 0
112	17– 4	86	13– 3	61	10– 9	36	8–11
111	17– 2	85	13– 1	60	10– 8	35	8–10
110	17– 0	84	12–11	59	10– 7	34	8– 9
109	16–10	83	12–10	58	10– 6	33	8– 8
108	16– 8	82	12– 8	57	10– 6	32	8– 7
107	16– 6	81	12– 7	56	10– 5	31	8– 6
106	16– 5	80	12– 6	55	10– 4	30	8– 5
105	16– 3	79	12– 4	54	10– 3	29	8– 4
104	16– 2	78	12– 3	53	10– 2	28	8– 3
103	16– 0	77	12– 2	52	10– 1	27	8– 2
102	15–11	76	12– 0	51	10– 0	26	8– 1
101	15– 9	75	11–11	50	9–11	25	8– 0
100	15– 8	74	11–10	49	9–11	24	7–11
99	15– 6	73	11– 9	48	9–10	23	7–10
98	15– 4	72	11– 8	47	9– 9	22	7– 8
97	15– 2	71	11– 7	46	9– 8	21	7– 6
96	15– 0	70	11– 6	45	9– 7	20	7– 5
95	14–10						

Age norms are particularly valuable where it is desirable to make use of the test results of individual pupils rather than of entire classes or grades. The following illustration of such age equivalents (Table 30) is based upon data accompanying the *New Stanford Achievement Tests.*[8] It is used here because age norms for high school tests are generally not available. On the basis of these age norms the achievement age of a pupil may be obtained without regard to his grade classification. For example, suppose a pupil 11 years and 3 months of age makes a composite score of 80 on the advanced examination. A reference to this table immediately reveals that fact that he has made a score which is typical of a large group of pupils 12 years and 6 months of age. This age, 12 years and 6 months, thus becomes the educational or achievement age of this pupil. In view of the fact that the pupil's actual age is only 11 years and 3 months, this means that his achievement is one year and three months above the typical score of pupils of his age. Under the heading of "Derived Scores" (page 231), the method of turning these age scores into the form of quotients is discussed.

125. **Interpretation of Grade and Age Norms.** The interpretation of results from diagnostic tests for both individual pupil and class diagnosis may also be made direct from grade and age norms. This is illustrated by data from the *Compass Diagnostic Tests in Arithmetic* (Table 31). Results from this elementary school test are used here for illustration because similar data are not available for high school tests. A fifth grade pupil R.H., age 11 years and 3 months, made the scores on the five parts of *Test* 7, *Multiplication of Fractions and Mixed Numbers,* shown in Table 31 on page 229.

The age equivalent of 10 years and 9 months means that a total score of 75 points on this test is the typical or average score made by pupils who were 10 years and 9 months of age. See Table 33.

The symbol H6 means that a score of 11 points on Part 3 of this test is equivalent to the typical score made by pupils in the advanced section of the 6th grade. See Table 32.

[8] *New Stanford Achievement Tests, Manual of Directions,* Kelley, T. L., Ruch, G. M., and Terman, L. M. World Book Company.

TABLE 31

PUPIL RECORD FROM COMPASS DIAGNOSTIC TEST NO. 7

Part..........	1	2	3	4	5	Total
Score..........	17	13	11	30	4	75
Age Equivalent..	.....	.....	.....	.....	.....	10 yrs. 9 mos.
Grade Equivalent	H6	H6	H6	−5	−5	

The symbol —5 means that a score of 30 points on Part 4 of this test is less than available norms for the 5th grade. See Table 32.

TABLE 32

GRADE NORMS FOR COMPASS DIAGNOSTIC TEST NO. 7

Part	Grades						
	H5	L6	H6	L7	H7	L8	H8
1	16	16	17	17	17	17	18
2	12	12	13	13	14	14	15
3	9	10	11	11	12	12	12
4	35	42	49	53	57	62	67
5	8	9	10	11	12	12	13
Total..	80	89	100	105	112	117	125

The age equivalent and grade equivalents which also appear in the table are obtained from the accompanying tables of norms for the test (Tables 32 and 33). For example, a score of 17 on Part I of Test 7 is the typical or normal score for pupils in the advanced section of the sixth grade. This pupil's achievement in Parts 2 and 3 is also equal to

TABLE 33

AGE-EQUIVALENTS OF SCORES
COMPASS DIAGNOSTIC TEST No. 7

Scores	Age Equivalents Years	Months
66–70	10	6
71–75	10	9
76–79	11	0
80–84	11	3
85–89	11	6
90–94	11	9
95–98	12	0
99–102	12	3
103–105	12	6
106–109	12	9
110–112	13	0
113–115	13	3
116–117	13	6
118–120	13	9
121–124	14	0
125–128	14	3
129	14	6

that of high sixth-grade pupils. However, in Parts 4 and 5 his achievement is considerably below that of the fifth grade norms. For purposes of individual diagnosis this is the significant information. This pupil is still in need of further practice on the multiplication of fractions. Possibly his lack of experience with exercises calling for the finding of errors may account for his relatively low score on that function.

A real understanding of the meaning and use of test norms in the interpretation of test scores can come only as a result of extensive first hand experience with their use. Problem 29 in the *Work-Book* shows the steps in the process of deriving a set of grade norms for a series of practice exercises in arithmetic. Problems 19, 20 and 21 on pages 244 and 245 of this volume and Problems 34 to 38 inclusive in the *Work-Book* give further practice on the summarizing and interpreting of test scores from various types of tests.

III. Types of Derived Scores

126. Function of Derived Scores. Test scores, which are used to describe the performance of pupils, are expressed in a variety of different units and in relation to a variety of different scales of measurement. In some tests the scale of measurement or the unit of measurement used may be relatively large. Pupil scores expressed in terms of these large units may be small. A long test composed of many exercises may result in large scores. Some common basis must be utilized if comparisons of scores based on these widely different types of scales are to be possible. This is illustrated in Table 34.

TABLE 34

COMPARISON OF UNITS OF MEASUREMENT USED IN TWO READING TESTS IN TERMS OF GRADE NORMS

(*Thorndike-McCall* and *Iowa High School* Reading)

Test	Grade			
	9	10	11	12
Thorndike-McCall Crude Score	13	17	19.5	21.5
Iowa H.S. Silent Reading Test Total Comprehension Score, Tests 1–5......	80	93	109	123

Table 34 presents a comparison of the grade norms for two silent reading tests both of which are standardized for use in grades 9 to 12. On the *Thorndike-McCall Silent Reading Scales* a crude score of 13 constitutes the normal seventh grade score. Contrast with this a score of 80 as normal for ninth grade pupils on the *Iowa Silent Reading Test, Advanced.* The comparison is even more striking when the normal scores for the twelfth grade achievement are compared. It is thus apparent that a given point score on one test may mean exceptional achievement while a simi-

lar point score on another test may mean exceedingly poor achievement for the grade. To repeat, if scores on different tests are to be compared directly, they must be reduced to a common denominator. That is, if a score of 62 points on one test is poor performance but on a second test because of a different system of scoring is very good performance, we must have some way of translating both of these scores of 62 points to a common system of scores so that their difference in meaning will be clear. A number of proposals have been made for the calculation of derived scores which will partially take care of this difficulty.

127. **Grade Norms as Derived Scores — Type 1.** Point scores yielded by a test may be expressed in terms of the grade norms for the test. This is a very effective method of changing a test score into form which is readily understood, and which at the same time expresses the level of achievement represented by the score. These grade equivalents can be found for scores on any test having reliable grade norms. Their significance is, of course, subject to the reliability of the test norms as well as to the reliability of the test itself. Following its somewhat general use the grade equivalent has been featured in connection with certain tests for which age norms were not available. These grade equivalents (G-Scores) are based directly upon the grade norms for the tests, and may be computed quite readily for any test. The method is illustrated in Table 35. This table shows, for example, that a point score of 15 is typical fourth grade achievement while a score of between 9 and 10 is typical seventh grade achievement. Each point value from 9-10 to 15 is assigned its proportionate position on the grade scale. A point score of 12 is considered as equivalent to the 3.5 grade, 14 to 3.8 grade, etc. In other words, a score of 12 is such a score as may be expected from pupils half-way through the third grade.

Grade scores of this type are easily constructed from the grade norms ordinarily accompanying standard tests.[9] The degree of fineness with which this scale may be developed depends upon the characteristics of the test scale itself. If the test scale consists of many fine units, the grade scores

[9] See Problem 30 of the *Work-Book in Educational Measurements.*

TABLE 35

G-SCORES FOR *Thorndike-McCall Silent Reading Scales*

Crude Score	G-Score	Crude Score	G-Score	Crude Score	G-Score
35	15.6	23	6.5	11	3.3
34	15.3	22	6.1	10	3.1
33	15.0	21	5.8	9	2.9
32	14.7	20	5.4	8	2.9
31	14.4	19	5.1	7	2.8
30	13.3	18	4.7	6	2.6
29	11.7	17	4.5	5	2.6
28	9.2	16	4.2	4	2.4
27	8.3	15	4.0	3	2.3
26	7.7	14	3.8	2	2.1
25	7.2	13	3.7	1	1.8
24	6.8	12	3.5	0	1.5

assigned can be correspondingly minute. For example, a test with a relatively coarse unit requiring only one or two such units to span the entire difference between two grades would, of course, permit the establishment of the grade scores on a similarly crude basis. On the whole there is probably an element of danger in the general use of G-scores due to the fact that they tend to foster a feeling of mathematical exactness which does not exist. It is very doubtful if there is a single test in existence for which the G-scores for an individual pupil have meaning. The significance of such scores is entirely dependent upon the reliability of the test and of the norms from which they are developed.

This method of expressing test scores is particularly useful where test results are used for purposes of gradation and classification of pupils. The chief convenience lies in the fact that it expresses the achievement of the pupil in terms of the grade unit of school classification, a practice which is followed much more frequently in the elementary school than it is in the high school. In fact, the numerous short-unit courses in the high school make it almost impossible to illustrate this procedure with data from high school tests.

The manner in which the test scores of a pupil on a number of standardized tests can be shown graphically by means of these grade equivalents is illustrated in Figure 13, page 236. Elementary school tests are used in this illustration for the reason that suitable data from high school tests are not available. Figure 13 is representative of a simple type of test record card. It not only reveals the achievement of the pupil on the tests, but it also reveals at a glance the variation of ability in the different tests, and the accuracy of his classification. By changing the test score into its equivalent grade-level[10] (G-scores) value a simple uniform graphic record of each pupil may be made. The following illustration will make this clear.

A pupil, Robert K., 11 years and 9 months of age, presents himself for entrance into a certain school system. An examination of his credentials shows that the work he has been doing in the school from which he has transferred does not parallel the work in this school system sufficiently to make his classification a simple matter. The administration of four or five well-standardized tests in achievement together with one or two reliable group mental tests will provide data of much value in placing this child in the new school system at a level at which he will be able to do his best work and which will be thoroughly fair to him. Assume that one mental test and an achievement test in reading, arithmetic, language, writing and spelling are used, and the following scores are earned:

Pupil: Robert K.

Test	Score
Mental Test	118
Reading	38
Arithmetic	25
Language	21
Writing (Quality)	57
Spelling	61

10 See *Glossary* for further definition of this term.

The grade norms for the tests are as follows:[11]

Grade	Mental Test	Reading	Arith-metic	Language	Writing	Spelling
4	81	....	13	....	50	32
5	96	23	21	10	55	47
6	109	32	28	16	59	57
7	120	40	32	21	64	66
8	129	46	35	25	70	74
9	136	50	36			

By referring the score made on any of these tests to the table of grade norms it is possible to determine the grade level of achievement nearest which the accomplishment of this child places him. For example, a score of 38 on the Reading Test places him three-fourths of the way between the norms for the sixth and seventh grades. Accordingly, this score may be thought of as 6.7 on the grade level scale. In a similar way the arithmetic score is four-sevenths of the distance between the norms for the fifth and sixth grades and this score may be considered as 5.7 points on this scale.

The graphic record card given on page 236 (Figure 13) presents the record of this student. At the time the tests were given he was in the 5A grade. This may be shown by a heavy vertical line under the 5th grade. On the mental test which was used he made a score almost equal to a seventh grade child; on reading he scored midway between the sixth and seventh grade norms. On all but two tests, arithmetic and writing, he exceeded sixth grade norms. Because of his superior mental ability and his excellent work in reading and language, and in spite of his somewhat lower scores in arithmetic and writing, it was thought advisable to give him a chance to attempt to carry 6A work. Many other similar uses for this type of graphic record may be worked out by the interested teacher.

128. **Age Scores as Derived Scores — Type 2.** One of the most widely used and most meaningful types of derived scores is the age score. This score is obtained by giving

[11] These norms are hypothetical throughout, and are merely used for illustration.

Graphic Record Card

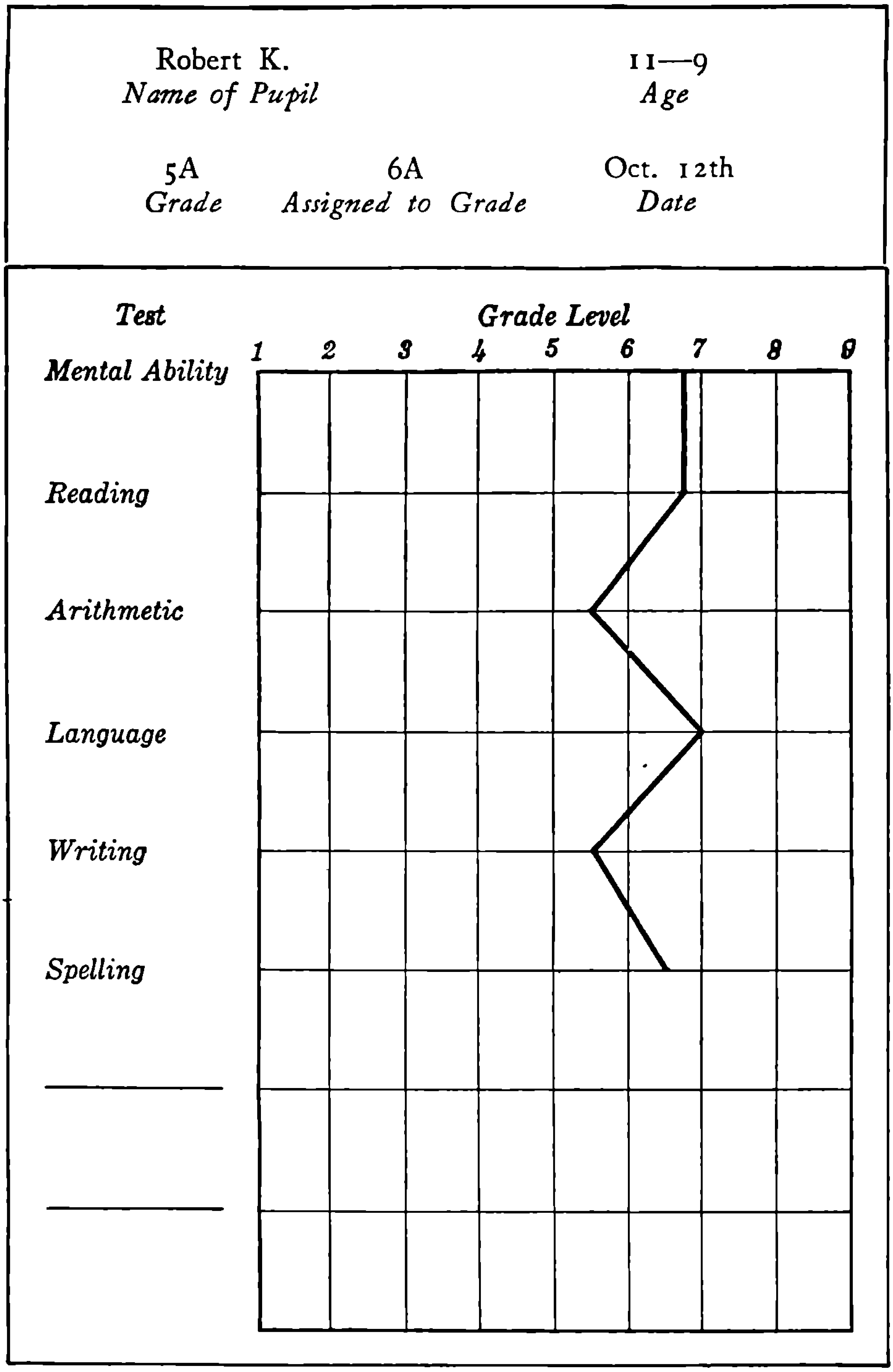


Figure 13. Pupil's Graphic Record

the test to all pupils of certain ages who are available for the purpose regardless of the grades in which they may be located. The test scores are then distributed by age groups and the median or average scores computed. The resulting scores for the various age groups are taken as expressing the typical achievement of these groups. For example, if

TABLE 36

C-SCORES AND READING AGES FOR UNIT SCALES OF ATTAINMENT[12]

C-scores	Reading Ages, Months	C-scores	Reading Ages, Months
91	198	75	150
90	195	74	147
89	192	73	144
88	189	72	141
87	186	71	138
86	183	70	135
85	180	69	132
84	177	68	130
83	174	66.5	128
82	171	65	126
81	168	63.5	124
80	165	62	122
79	162	60	120
78	159	58.5	118
77	156	57	116
76	153	55	114

it is found that typical children ten years and three months of age make an average score of 46 points on a certain test, then any other child making a score of 46 points on the test is said to have a test age of 10 years and three months. If the test is a mental test the resulting age score is a mental age. If it is an achievement test the resulting age is an achievement age, the subject age assigned depending upon the subject tested. The age score is fundamental to the development of quotients in the field of measurement.

[12] Adapted from *Unit Scales of Attainment* Profile Chart, and reproduced with the permission of the publishers, Educational Test Bureau, Minneapolis and Philadelphia, 1935.

The use of age norms as a means of giving meaning to test scores is illustrated in Table 36, which is adapted from the profile chart for the *Unit Scales of Attainment.*[13] This table shows the C-scores and Reading Ages corresponding to a given number of questions answered correctly by pupils taking the test. The Reading Ages of the pupils discussed in Table 37 were obtained by referring to Table 36. This means that Pupil A's score in the number of questions answered correctly is equivalent to a C-score[14] of 68, and a reading age of 130 months. Whether or not this is a good score for Pupil A depends on how this reading age compares with his mental age and chronological age.

TABLE 37

DATA FOR READING AGES

Pupil	Questions Answered Right	C-score	Reading Age, (Months)
A	19	68	130
B	34	89	192
C	26	60	120

The use of age scores in relation to the chronological age of the pupil is presented in Section 130 under the heading of Quotients as Derived Scores. Problem 30 of the *Work-Book in Educational Measurements* illustrates more completely the method of determining age scores from tables of age norms for tests.

129. Percentile Ranks as Derived Scores—Type 3. The test performance of a pupil may be expressed in terms of the position which his score occupies in the standard distribution of scores for his particular grade. This is accomplished by

[13] *Unit Scales of Attainment,* published by Educational Test Bureau, Minneapolis and Philadelphia, 1933.

[14] The C-scale unit is defined by Van Wagenen in his manual for the Unit Scales of Attainment as "approximately one-tenth of the quartile of all pupils of the same chronological age." . . . It is called C-score because by its use it is possible to see "how difficult tasks a pupil is able to attempt and get half of them right, just as in the case of mental ages on the Stanford and the Kuhlmann revisions of the Binet-Simon Individual Intelligence Examinations."

dividing the distribution so that exact divisions contain the same per cent of the total number of cases. One plan is to divide the distribution into quarters (quartiles), another into tenths, twenty-fifths, hundredths, etc. When the distribution is divided into one hundred parts, each part is called a *percentile.* (See Section 112, Chapter VIII.) Ability as represented by a test score is assigned as a position in the rank of one hundred individuals measured by the test. Percentile ranks are particularly valuable where no standards are available, as in the case of most informal objective examinations, or where standards are limited in meaning as is the case with most high school tests. The percentile rank[15] also affords a very convenient means of reducing to a comparable basis measures from tests of other school subjects which are otherwise not comparable.

The methods of actually turning test scores into percentile scores are discussed in detail in Sections 110 to 113 of Chapter VIII. Problems 27 and 28 of the *Work-Book in Educational Measurements* also give additional practical uses of percentile ranks as derived scores.

130. **Quotients as Derived Scores — Type 4.** Age norms for tests permit the expression of a relationship between the chronological age and the mental and subject ages of pupils. The relationship thus expressed is called a *quotient.*

Intelligence Quotients: When mental age is used as the numerator in the fraction the resulting quotient is called an *Intelligence Quotient.* The following example will illustrate the method of determining the Intelligence Quotient (*I.Q.*). Suppose a pupil exactly eleven years of age earns a score on a mental test which gives a mental age of twelve years and ten months. The *I.Q.* is determined as follows:

$$\frac{\text{Mental Age (in months)}}{\text{Chronological Age (in months)}} = \frac{154 \text{ mos.}}{132 \text{ mos.}} = 1.17 = 117$$

In practice the result of dividing the mental age in months by the chronological age in months is multiplied by 100 to permit the expression of the quotient as a whole number.

[15] For a more complete discussion of percentile ranks and the method of computing them see Chapter VIII, pages 203 to 206.

An *I.Q.*, of 100 means perfectly normal development on the part of the individual, within the limits of the reliability of the test, of course. Average *I.Q.*'s range from 90 to 110. *The I.Q. is an index of relative mental development and in no sense expresses absolute mental ability. Its use diregards the level of mental ability, but expresses more exactly the relative brightness of the pupil.* On this account it becomes a valuable instrument for the sectioning of pupils after general classification has been effected by means of mental and educational age data.

Educational Quotients: Many achievement tests permit the determination of quotients in a manner similar to that followed in the case of mental tests. Scores from a series or battery of achievement tests such as the *New Stanford Achievement Tests*,[16] the *Unit Scales of Attainment*,[17] the *Metropolitan Achievement Tests*,[18] or the *Modern School Achievement Tests*,[19] for example, afford a very valuable bit of supervisory information when the educational ages and the educational quotients of the various pupils are determined.

The method of turning an educational test score into a quotient representative of the relative development of the individual pupil on the test used may be illustrated by referring to Table 36 on page 237. This table gives the C-score and Reading Age equivalents for the reading section of the *Unit Scales of Attainment.* In connection with the illustration on page 238 the scores of three pupils on this test were used. (Table 37.) The same pupils are utilized here for the purpose of showing how the Reading Quotients are computed. Table 38 summarizes all necessary data on these pupils.

The score of Pupil A gives him a reading age of 130 months according to Table 36, page 237. He is found to be 160 months of age chronologically. He is therefore not developed in his reading ability as high as might be expected

[16] *New Stanford Achievement Tests.* Kelley, Ruch, and Terman. World Book Company, 1929.

[17] *Unit Scales of Attainment.* Van Wagenen, et al. Educational Test Bureau, Minneapolis, Philadelphia, 1933.

[18] *Metropolitan Achievement Tests.* World Book Company, 1933.

[19] *Modern School Achievement Tests.* Bureau of Publications, Teachers College, Columbia University, New York, 1933.

from his chronological age. His reading quotient is therefore less than normal, or is equal to $\frac{130}{160}$, or 81. The reading quotients for the other pupils in this illustration are found in a similar manner.

TABLE 38

DATA FOR READING QUOTIENTS

Pupil	Questions Answered Right	C-scores	Reading Age (Months)	Chronological Age (Months)	Reading Quotient
A	19	68	130	160	81
B	34	89	192	181	106
C	26	60	120	96	125

Accomplishment Quotients: Teachers often raise the issue that a certain pupil is not making the progress in his school work that he is capable of making. When both intelligence and subject tests are given and the scores are turned into mental and subject ages it is possible to answer this question directly. The process is called finding the Accomplishment Ratio or the Achievement Quotient. The educational significance of this technique has been most completely presented and defended by Franzen.[20] The *Accomplishment Quotient* or *A.Q.* is nothing more than the ratio between the relative educational development of a pupil (*E.Q.*) and his relative brightness (*I.Q.*). In symbols this statement takes the form:

$$A.Q. = \frac{\frac{E.A.}{C.A.}}{\frac{M.A.}{C.A.}} = \frac{E.A.}{M.A.}$$

where *E.A.* is the educational age of the subject and *M.A.* is the Binet mental age of the subject. Since chronological

[20] Franzen, R., "The Accomplishment Ratio; a Treatment of the Inherited Disparity in School Product." Teachers College, Columbia University Contributions to Education No. 125; New York, 1922.

age (*C.A.*) is common to both factors in the ratio the *A.Q.* may be found by dividing *E.A.* directly by *M.A.* If a pupil's achievement (*E.A.*) is in keeping with his power to learn (*M.A.*) his *A.Q.* is 100. Obviously, if the quotient is below 100 the pupil is not working up to capacity. On the other hand a study[21] of highly motivated instructional drives on certain subject matter tends to show that an *A.Q.* of more than 100 is obtainable. It is probable, however, that in the majority of cases, A.Q.'s of over 100 will be the result of norms on subject matter tests which are spuriously low. When we have true or correct norms (representative of what is done by normal children under approximately perfect conditions) it seems certain that A.Q.'s of more than 100 will disappear. In fact it may not be unlikely that the frequency with which A.Q.'s of over 100 are obtained may be used to demonstrate how greatly many achievement test norms fail to represent the conditions which norms and standards of achievement imply.

Table 39

DATA FOR ACCOMPLISHMENT QUOTIENTS

Pupil	Chronological Age (Months)	Mental Age	I. Q.	Reading Age	Accomplishment Quotient
A	160	136	85	130	96
B	181	216	119	192	89
C	96	133	139	120	90

The method of computing the accomplishment quotient may be illustrated if we utilize the records of Pupils A, B, and C, from Tables 37 and 38 on pages 238 and 241. The reading ages of these pupils are available. Let us assume that mental tests have been given and the mental ages recorded in Table 39 are found. This table summarizes all data necessary for the calculation of the Reading Accom-

[21] Lessenger, W. E., "Motivation and the Accomplishment Quotient Technique." University of Iowa Study in Education, Iowa City, 1925.

plishment Quotients for these pupils. The I.Q. is the ratio of the mental age and the chronological age of the pupil. The I.Q. for Pupil A is $\frac{136}{160}$, or 85. The Reading Accomplishment Quotient is the ratio of the reading age of the pupil to his mental age. For Pupil B this ratio is $\frac{192}{216}$, or 89. In theory an accomplishment quotient of more than 100 is an impossibility, since it is maintained that a pupil cannot be expected to develop a special skill higher than his general mental ability. The probabilities are that when such quotients exceed 100 the reason for it may be traced to the unreliability of the tests themselves. The Accomplishment Quotient of Pupil A, 96, means that his reading ability as measured by the test used is developed practically up to the level of his mental ability.

Further experience in the finding of quotients from actual school data of the different types is given in Problem 21, at the end of this chapter, and in Problems 30 to 33 in the *Work-Book in Educational Measurements.*

IV. Summary

This chapter presents the next logical steps in the interpretation of test results after the necessary statistical summaries have been made. These were discussed in Chapter VIII. The meaning of norms and standards, their use in giving meaning to test scores, the reasons for the need for age norms for tests, and the practical uses which may be made of derived scores are also presented and illustrated in this chapter.

The problems involved in the interpretation of the results of testing are many and important. The treatment accorded to them in this volume is so limited by the requirements of space that the attention of the student interested in securing a more critical basis for the interpretation of test results is called particularly to the references given at the end of this chapter.

Problem-Exercises

Problem 19

Comparing Test Scores with Grade Norms

Use the table of grade levels adapted from the norms for the Thorndike-McCall Silent Reading Scales, Table 34, and complete the work begun in this list of pupil scores. The first two are filled in for illustration.

Pupil	Score	Grade Level	Pupil	Score	Grade Level
1	4	2.4	7	8	
2	26	7.7	8	18	
3	6		9	28	
4	30		10	12	
5	15		11	22	
6	1		12	14	

Problem 20

Finding Age Equivalents from Test Scores

Use the table of age equivalents for the *New Stanford Achievement Tests*, Table 30, page 227, and complete the work begun in this list of pupil scores. The first two are filled in for illustration.

Pupil No.	Composite Scores	Age Equivalent	Pupil No.	Composite Scores	Age Equivalent
		Yr. Mo.			Yr. Mo.
1	25	8 0	9	79	
2	34	8 9	10	66	
3	20		11	30	
4	65		12	44	
5	51		13	82	
6	95		14	90	
7	106		15	70	
8	112		16	48	

Problem 21

Finding Quotients from Ages Given

Complete the work begun in the accompanying list of pupils and scores. Use the accompanying table of age norms for the *Haggerty Intelligence Examination, Delta* 1, to determine age equivalents. The first two quotients have been correctly computed for illustration.

REVISED AGE NORMS FOR
INTELLIGENCE EXAMINATION: DELTA I

Year	Month											
	0	1	2	3	4	5	6	7	8	9	10	11
6	30	32	34	36	38	40	42	44	46	48	50	52
7	54	55	56	57	57	58	59	59	60	61	62	63
8	64	65	66	66	67	68	68	69	70	70	71	71
9	72	73	74	74	75	76	76	77	78	78	79	79
10	80											

Pupil	Life Age		Point Score	Age Equivalent		Quotient
	Yrs.	Mos.		Yrs.	Mos.	
1	7	0	54	7	0	100
2	7	2	59	7	6	105
3	6	6	63			
4	8	5	58			
5	7	6	55			
6	7	0	36			
7	7	4	55			
8	8	1	42			

Selected References

Buckingham, B. R., *Research for Teachers.* New York: Silver, Burdett and Company, 1926.

Franzen, Raymond, *The Accomplishment Quotient Technique.* New York: Teachers College Contributions to Education, No. 125. Columbia University, 1922.

Greene, H. A., *Work-Book in Educational Measurements.* New York : Longmans, Green and Company, 1930.

Kelley, T. L. ; Ruch, G. M. ; Terman, L. M., *The Stanford Achievement Tests.* Yonkers, New York : World Book Company, 1929.

Kelley, T. L., *Interpretation of Educational Measurements.* Yonkers, New York : World Book Company, 1927.

Odell, C. W., *Educational Measurement in High School.* New York : The Century Company, 1930.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago : Scott, Foresman and Company, 1929.

Ruch, G. M .; Knight, F. B. ; Greene, H. A. ; Studebaker, J. W., *The Compass Diagnostic Tests in Arithmetic.* Chicago : Scott, Foresman and Company.

Ruch, G. M., and Stoddard, G. D., *Tests and Measurements in High School Instruction.* Yonkers, New York : World Book Company, 1927. Chapter XIX.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York : Silver, Burdett and Company, 1928.

Trabue, M. R., *Measuring Results in Education.* New York : American Book Company, 1925.

Wilson, G. M., and Hoke, K. J., *How to Measure (Revised).* New York : The Macmillan Company, 1928.

Woody, Clifford, and Sangren, Paul V., *Administration of the Testing Program.* Yonkers, New York : World Book Company, 1932.

CHAPTER X

DIAGNOSIS, ANALYSIS AND REMEDIAL TEACHING IN THE HIGH SCHOOL

THIS chapter treats the following diagnostic functions of tests and their use in remedial teaching:

a. Meaning of diagnosis in education.
b. Teacher responsibility for the diagnostic use of tests.
c. Exact diagnosis by means of education tests.
d. Relation of remedial teaching to exact diagnosis.
e. Principles for the construction of remedial materials.
f. Concrete application of these principles to the construction and use of remedial materials.

I. THE PLACE OF DIAGNOSIS AND ANALYSIS

131. **The Meaning of Diagnosis.** Educational diagnosis implies the use of procedures, more or less technical in character, designed to locate specific learning and instructional difficulties and if possible to determine their causes. For the medical expert diagnosis means the careful and extensive observation of the patient under controlled conditions. It includes the use of professional instruments such as the clinical thermometer, the stethoscope, the microscope, etc., which make observations more exact and more objective. It means the assembly of a complete case history of the background of the difficulty leading up to the present physical crisis. It means the examination and analysis of many similar cases in order that common factors may be identified. For the teacher, diagnosis has many of the same implications, but unfortunately much of the exactness, objectivity, precision of the medical diagnostician's instruments appears to be missing in the teacher's equipment. Only recently have objective measuring instruments, capable of reasonably

exact diagnosis, become available to the pedagogical diagnostician. The well-prepared modern teacher now has at hand a refined statistical technique, exact and analytical diagnostic tests in different subjects, diagnostic charts, devices for measuring aural acuity, instruments for measuring visual acuity, eyedness, muscular imbalance in the eyes, binocular vision, binocular fusion, and many other highly important qualities which may account for a pupil's lack of progress in many fields of learning. It is thus apparent that diagnosis in education is rapidly becoming accurate and scientific.

132. **Diagnosis in Education.** While diagnosis through measurement represents only a single phase of the work in education, it is nevertheless an exceedingly important aspect of it. The analysis and diagnosis of difficulties underlying educational accomplishments undoubtedly constitute the high point in the supervisory and instructional uses of educational tests. General impressions of strengths or weaknesses are supported or denied by test results. Deficiencies of a general nature are revealed and brought to the light by general survey tests. Specific weaknesses, and to a certain extent causes of weaknesses, are identified by the use of properly selected diagnostic tests. Practically all of the more exact types of diagnostic procedures, such as the location of defects in speech, hearing, vision, are dependent upon educational test results for their initial steps. These points will be discussed much more thoroughly in the later chapters on reading and language.

133. **Analysis as the Basis of Diagnosis.** The successful development of the many sets of habits which constitute the bulk of our school learning depends upon the care with which the underlying and basic habits of the subjects themselves are recognized and utilized in the teaching. If, in arithmetic, a habit of responding correctly to the 100 basic addition facts is essential to success in addition, then this fact must be recognized and stressed. If it can be shown that in addition to these basic skills there are other related but somewhat higher skills which are fundamental to successful work these also must be recognized and developed. If it can be shown that teaching a child to add consists not only in

the mastery of the 100 basic combinations but also involves high levels of skill such as knowledge of the higher decade addition facts, the bridging of the tens, the control of the attention span, and carrying from one column to the next, the teacher's task is made much more obvious. Similarly, if it can be shown that silent reading comprehension is not a single isolated ability but a composite of many elements such as knowledge of word meanings, ability to get meaning from sentences, ability to arrange thought units and sentence units into logically organized wholes, ability to find desired material quickly, etc., the teacher has a real basis for her instructional procedures. Language is another basic subject in which many delicately balanced skills are interwoven in an extremely complex manner. Here again the elements of achievement in the total process must be identified. Blind trust in general practice on the total skill must necessarily give way to the exact identification and discovery of the particular points of pupil weakness as a basis for special emphasis.

Good diagnosis must parallel exactly the processes of good teaching. Effective diagnostic materials in any school subject can be prepared only after the skills contributing to success in this field have been isolated and identified. Psychologically the reason for this is that on the whole the child learns to do what he practices and not something else. Remedial work, accordingly, can function only when the exact level at which pupil mastery breaks down has been located. Thus the analysis must be penetrating, and the diagnosis must be exact.

134. **Specific Nature of Diagnosis.** Diagnosis must be more exact than broad statements of general functions. It is not enough to discover that a child is unable to read silently. The exact nature of his handicap must be revealed before a remedial program[1] can be undertaken. The more specific the diagnostic information revealed the more exactly the remedial material can be made to fit the need. To re-

[1] The one exception to this is when the child is weak in every skill constituting reading ability. In this case he may not need remedial drill on certain skills, or even complete re-teaching of the total process, but rather an analysis of certain physical or psychological difficulties in his visual or perceptual equipment which is basic to learning to read.

turn to a frequently used illustration, it is found by diagnosis that the child is unable to add, but unless diagnosis can determine the exact point at which his mastery of addition breaks down, teaching or remedial efforts are largely wasted. One of the outstanding reasons why more effective teaching and remedial work has not been done in certain fields is that no adequate analysis of basic skills has been made. Concrete illustrations of this need are given in connection with a related discussion in Section 136 of this chapter.

135. **Importance of the Diagnostic Interpretation of Test Results.** There is nothing mysterious in the administration of a standard or informal objective test which produces an immediate improvement of instruction in the classroom. Although tests have no inherent power to improve instruction directly, they bring it about indirectly by revealing existing conditions. Remedial teaching is the result of deliberate constructive effort by the teacher after the particular points of weakness in the instruction of the pupils have been revealed by the tests. The ease, clearness, and directness with which these needs are revealed by the test are measures of its real educational value. Too few of existing standard tests are so constructed as to permit the interpretation of their results directly in terms of an effective remedial procedure. However, this seems to be no good reason for the failure of teachers to apply more directly the results of this work in testing to the improvement of their teaching practice. Just as the data revealed by the navigator's instruments require calculation and interpretation, so is it necessary to analyze carefully test data in order to make it the basis of a genuine remedial program. Nor is the teacher the only one to profit from a frank study of test scores. Pupils themselves will often take an active part in remedial enterprises when they are given the facts concerning their own performances on standard tests.

The interpretation of test scores and the planning of remedial procedures are the most difficult parts of the use of standard educational tests. Moreover, they are most important parts. One of the greatest needs in education today is the provision for genuine diagnostic testing in all instructional fields, supplemented by valid remedial work

designed to correct the weaknesses and defects of individual pupils as revealed by the tests. Makers and users of tests have been for the most part content merely to measure. It is coming to be recognized that the giving of a test with a study of the resulting scores will not in itself improve the classroom situation. This is especially true if the test is of the general survey type. It is interesting to learn as a result of using tests in the classroom that an individual pupil or the entire class is below standard in the subject, but unless we learn with some exactness what causes the low level of achievement our testing program will do little if anything more than supply us with interesting information. Teachers and supervisors have a right to expect that something more constructive will be provided in exchange for the time required for classroom testing.

II. The Place of Remedial Instruction

136. Exact Diagnosis the Basis for Remedial Work. Exact diagnosis of class and individual pupil difficulties coupled with the application of specific remedy is the heart of enlightened use of exact methods of teaching. The success of the remedial or corrective teaching depends upon the accuracy and detail with which the specific skills involved in the successful achievement in the subject are identified and isolated in the test. General survey tests can not supply this information in sufficient detail. This is illustrated in the case of the record of a single pupil given in Table 40.

This table shows the test scores of Pupil H.J.A., a ninth grade pupil in School R, on four tests, two in first year algebra, and two in language. Two of the tests are analytical or partially diagnostic while the other two are excellent examples of general survey tests. The tests used in this illustration are the *Columbia Research Bureau Algebra Test, Test* 1 [2], *Douglas Standardized Diagnostic Test in Elementary Algebra, Test* 1 [3], the *Wilson Language Error Test, Test* 1 [2], and the *Iowa Elementary Language Test* [4].

[2] World Book Company, Yonkers, New York.
[3] C. A. Gregory Company, 345 Calhoun Street, Cincinnati, Ohio.
[4] Educational Test Bureau, Inc., Minneapolis, Minnesota.

TABLE 40

RECORD OF TEST SCORES OF PUPIL H.J.A.

(Grade 9A; School R)

Columbia Research Bureau Algebra Test	Pupil Scores	End-of-year Norms for Grade
Part 1	14	11.5[5]
Part 2	3	5.0[5]
Douglas Standard Diagnostic First Year Algebra Test		
Test 1 Fractions	5	2.4
Test 2 Factoring	1	4.1
Test 3 Formulas and fractional equations	4	3.1
Test 4 Simultaneous equations	4	3.6
Test 5 Graphical representation and interpretation	4	2.5
Test 6 Square roots, exponents, and radicals	2	2.7
Test 7 Quadratic equations	2	3.4
Total	22	21.8

	Pupil Scores		End-of-year Norm for Grade	
	Parts	Total	Parts	Total
Wilson Language Error Test, Test 1		22		23
Iowa Elementary Language Test				
Test 1 Word Meaning		55		51
Synonyms	28		24	
Opposites	27		27	
Test 2 Language Usage		58		58
Correct Word	22		23	
Correct Word Form	22		23	
Faulty Expressions	14		12	
Test 3 Grammatical Form		10		16
Test 4 Sentence Sense		18		25
Test 5 Sentence Structure		12		18
Test 6 Capitalization-Punctuation		35		46
Capitalization	23		28	
Punctuation	12		18	
Test 7 Paragraph Organization		16		20
Total		204		234

[5] Estimated from data in Examiner's Manual. World Book Co., 1927.

The pupil's relatively high score of 14 points on Part I of the *Columbia Research Bureau Algebra Test* indicates an adequate mastery of the mechanics of algebra. The low score of 3 points on Part II of the test indicates that he is having some type of difficulty in problem solving in algebra. An examination of his paper shows that he attempted only seven of the problems and missed four. Furthermore not one of the problems he solved correctly called for the use of any factoring. A glance at his work on Part I, the section on the mechanics of algebra, at once makes his difficulty apparent. Of the eighteen examples he attempted, he solved fourteen correctly, but no one of the fourteen examples involved any difficult factoring. Thus it is apparent that his difficulty lies somewhere in the field of factoring.

The teacher at once gives him an analytical test, such as the *Douglas Standardized Diagnostic Test in Elementary Algebra.* His scores on the seven parts of this test are enlightening, as is shown in Table 40. Of the thirty-five opportunities provided in the seven tests to react to exercises in algebra, this pupil responded correctly to twenty-two. Since the sum of the norms provided for each of the seven tests is 21.8, the pupil is, in general, slightly above the norm. However, an examination of his record shows that his work with fractions is satisfactory, but that he fails to solve any of the more difficult examples involving factoring. In a similar way he is shown by Tests III, IV, and V to have adequate mastery in the skills covered by each of these test parts. According to these results, his mastery of square root and exponents as well as quadratic equations is not adequate. His weakness in these skills unquestionably accounts for his slow and inaccurate work in many of the examples and problems presented in the more general survey test.

In the further diagnosis of this pupil's school achievement two language tests were also used. The first, the *Wilson Language Error Test* is a general survey of language usage. On this type of test the pupil made a score of 22 points, which is only slightly below the norm for his grade. The administration of a more detailed analytical type of test, the *Iowa Elementary Language Test,* makes it very clear that while Pupil H.J.A. is very well equipped in word meanings

and language usage he is well below expected levels of achievement for his grade in such abilities as capitalization and punctuation. His recognition of sentences and his ability to arrange them into logically organized paragraphs are both somewhat below the grade norm. Here again is an instructional situation which is not at all revealed by the general survey instrument.

137. **Diagnosis as the Basis for Preventive Work.** An examination of the number and types of skills identified as a result of the diagnostic methods discussed in the preceding section leads to a suggestion of a still more constructive use of analytical and diagnostic test results. Diagnosis, as applied in education, has taken on a meaning indicative of a breakdown in method, a failure of instructional techniques to function. Unquestionably one of the basic purposes of diagnosis is the location of weaknesses and the determination of their causes, but there is nothing in the method which precludes its use in the prevention of weaknesses through anticipation of their causes. Out of the knowledge gained through the use of diagnostic procedures should come the basis for preventive work of all types. It is quite noticeable that the major emphasis in the fields of dentistry and medicine is not on correction but on prevention. The existence of a weakness in learning implies a failure at some point in the program. The discovery of it should not be marked as important merely because it is then possible to correct it. The real importance in the discovery should lie rather in the prevention of its reappearance elsewhere under similar conditions.

Another illustration from the field of medicine may make this point somewhat more concrete. In every medical examination for diagnostic purposes a complete analysis is made and an exact case record of all observations is kept. Out of the analysis of these records has come a better understanding of the causes and characteristics of certain types of human ailments. Out of this same type of analysis has also come the basis for much of the marvelous preventive work which characterizes modern medical science. In a similar way, accurate and detailed educational diagnosis may ultimately offer the basis for the development of a program

of preventive work in education. For example, if, after diagnosing the addition of fractions in the fifth grade the failure of pupils to reduce fractions in the answers is revealed as a common weakness, the obvious thing to do is to correct the defects at once and then proceed to reconstruct the first instruction so that next year the causes for this particular weakness may not operate so powerfully. In a similar way, any weakness identified now should afford the basis for decisions calculated to reduce the probability of their recurrence in the future.

It may be worth while to note that material so constructed as to be effective for remedial purposes is also sound to use for initial instruction. In fact, in most respects the chief distinction between good subject-matter content for initial teaching purposes and remedial drill purposes lies in *when* they are to be used. The most effective remedial drill for the pupil who does not have an adequate sight-meaning vocabulary for silent reading purposes is drill on the vocabulary he should have learned in the first place.

138. Necessity for Valid Drill for Each Identified Skill. Drill material of established validity must be provided for each specific skill which conditions achievement in the subject, if remedial work is to be effective. The validity of drill material depends to a large degree upon the accuracy and completeness with which the analysis of skills is made. Difficulties in subject-matter units which can be only hazily identified can not be remedied except by mere good luck. Drills must closely parallel the skills which they are supposed to remedy. If mastery of a certain minimal vocabulary is essential to effective silent reading comprehension, then drill on those particular words which constitute special weakness should preclude all other drill.

Perfect validity of drill material can be achieved only by taking a 100% sampling of all of the possible basic facts or skills in the particular field. Naturally this is impossible in certain cases, but it is nevertheless often possible to take such a large sampling that all of the most frequently used facts are included. Subject-matter fields vary widely in the ways in which they lend themselves to sampling of this kind. In fields such as reading or language a perfect

sampling is almost impossible to obtain. On the other hand many of the basic facts in arithmetic are so readily identified that they may be sampled 100% without difficulty. For example, a drill card for remedial (or initial instructional) purposes on the basic addition facts may sample the one hundred basic facts, all of the basic addition facts there are. On the other hand, there are so many higher decade addition facts (many of which have relatively slight social utility) that only a limited sampling may be taken. In practice it seems likely that a sampling of 450 such higher decade addition facts will cover most of these facts which appear in the problems of daily life.

Drill material designed for remedial and corrective use strikes directly at the heart of the trouble and wastes no time on skills which need no practice. Remedial drill material in which careful control is kept over the distribution of practice on the basic skills is almost certain to be more effective than random exercises, even assuming in both cases that suitable motivation for improvement is provided. The drill will be most productive which most nearly provides a complete coverage of the skills of basic importance in the hierarchy of habits upon which successful achievement in the subject depends. Poorly organized drill may or may not deal with all possible weaknesses, but it is almost certain to waste time on skills which are not in need of drill. The validity of the drill depends upon the degree to which this sampling covers the basic or fundamental skills, and the degree to which the exercises themselves develop the skills they purport to develop. There are a number of places in which this complex chain may break. The task of diagnostic and remedial treatment is to locate and repair quickly those links of the chain which have snapped under stress, or have rusted out through lack of use.

III. Summary

This chapter is in a very real sense the point at which all of the foregoing discussion of the meaning and uses of tests as instruments for the improvement of classroom instruction has been aimed. The chapter should clinch in the mind of

the teacher and student the often repeated fact that the underlying purpose of all testing is the accurate determination of class and individual pupil difficulties to the end that remedial instruction may follow. Of equal or even greater importance is the new angle which more exact analytical and diagnostic testing gives to the preventive phases of instruction through the anticipation of causes of weakness or difficulty.

Discussion Exercises

1. Expand the idea of the parallel between diagnosis and remedy in medicine and in education.
2. What reasons can you advance for the failure to develop adequate diagnostic and remedial materials in all subject-matter fields?
3. What are the essential differences between *remedial* and *preventive* work in education?
4. Show how a critical analysis of subject matter is necessary to the development of diagnostic tests.
5. What are the essential differences between good *diagnostic* material and good *remedial* material?
6. Select a school subject and show how the basic skills may be identified (diagnosed) in a way similar to that suggested in the discussion in your textbook.
7. To what extent are time and achievement standards fundamental to drill and corrective instruction in school subjects? Defend your position. (This may not be answered specifically in this chapter. It is food for thought.)
8. In a school field which you are likely to teach (your major or an important minor) suggest a number of specific skills which enter into successful work in this field, and parallel this with suggestions for remedial treatment.

Selected References

Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching*. Boston: Houghton Mifflin Company, 1931.

Brueckner, L. J., *Diagnostic and Remedial Teaching in Arithmetic*. Philadelphia: John C. Winston Company, 1930.

Buswell, G. T. (with L. John), *Diagnostic Studies in Arithmetic*, Supplementary Educational Monograph, No. 30. Chicago: University of Chicago Press, 1926.

Gray, W. S., *Remedial Cases in Reading; Their Diagnosis and Treatment.* Supplementary Educational Monograph, No. 22. Chicago: University of Chicago Press, 1922.
Greene, H. A., Ballenger, H. L., *Iowa Elementary Language Tests, Examiner's Manual.* Minneapolis: Educational Test Bureau, Inc., 1929.
Greene, H. A., Kelley, V. H., *Iowa Silent Reading Tests, Elementary.* Yonkers, New York: World Book Company, 1933.
Greene, H. A.; Studebaker, J. W.; Knight, F. B.; Ruch, G. M., *Economy Problem Solving Exercises.* Chicago: Scott, Foresman and Company, 1928.
Greene, H. A.; Studebaker, J. W.; Knight, F. B.; Ruch, G. M., *Economy Remedial Exercises in Arithmetic, Manual for Whole Number Series.* Chicago: Scott, Foresman and Company, 1927.
Sangren, P. V., "Methods of Diagnosis in Reading." *Elementary English Review,* Vol. 7: 105, April, 1930.

CHAPTER XI

MENTAL TESTS AS DIAGNOSTIC AIDS

THIS chapter treats the following aspects of mental measurement:

a. Meaning of mental ability.
b. Measurement of mental ability.
c. Kinds of mental tests.
d. Uses of mental tests.
e. Inter-dependence of achievement and mental test results.

I. THE MEANING OF MENTAL ABILITY[1]

139. **The Nature of Mental Ability.** The exact nature of the combination of abilities known as intelligence or mental ability is not well understood. However, we know enough about mental ability to be quite certain that individuals vary widely in the amount or quality of it they possess, and that within limits it can be measured. Authorities in the field of mental measurement are far from agreement as to what the term intelligence implies.[2] Some consider that intelligence

[1] Many people question the desirability of using the term *intelligence* when referring to tests such as are discussed in this chapter. The objections appear to be based mainly on the lack of definiteness of the term. Some would rather refer to such tests as Tests of Mental Ability, or simply as Mental Tests. The preference of the authors would be to discard both terms and designate all such measures as tests of *General Ability*.

[2] Some of the briefer definitions of mental ability are as follows:

(1) *Ballard,* "Intelligence is the relative general efficiency of minds measured under similar conditions of knowledge, interest and habituation."
(2) *Binet,* "Intelligence is judgment, or common sense, initiative, the ability to adapt oneself."
(3) *Burt,* "Voluntary attention is the essential factor of general intelligence."
(4) *Buckingham,* "Intelligence is ability to learn."
(5) *Colvin,* "An individual possesses intelligence insofar as he has learned, or can learn to adjust himself to his environment."
(6) *Henmon,* "Intelligence is capacity for knowledge and the knowledge possessed."
(7) *Gates,* "Intelligence is a composite measure of abilities to learn."
(8) *Knight,* "Intelligence is the ability to profit from experience."
(9) *Peterson,* "A biological mechanism by which the effects of a complexity

is best indicated by the ability of the individual to think, to solve problems, to adapt himself to new situations, to learn. Others hold that the abilities to perceive with speed and accuracy, to associate symbols, to give voluntary attention, to manipulate abstract concepts, and to reason are the best evidences of intelligence. Facility in the use of language itself is considered by some to be one of the very significant evidences of intelligence. For the purposes of this discussion *intelligence is the capacity for learning, plus the useful informations, skills, and attitudes which the individual has gained from reacting to his environment.* This rather liberal conception of intelligence permits it to fit into the educational program and also places in an acceptable light the majority of devices for the measurement of general mental ability.

II. The Measurement of General Mental Ability

140. **The Meaning of a Mental Test Score.** Interesting as is a consideration of these conflicting notions of the nature of mental ability, the question of what the mental test measures is probably of more vital concern to the classroom teacher. In a previous discussion of the meaning of test scores, attention has been called to the necessity for a clear understanding of such concepts as capacity, ability, training, and performance as the basis for defining the meaning of a test score. Courtis[3] has defined these terms in a clear and concise manner. He defines *capacity* as "the potential power to respond or potential development" ; *ability,* as "the power of the inherited and modified mechanism to respond to stimuli" ; *training,* as "experiences organized to produce the development or modification of mechanism or ability" ; and *performance,* as "the result of ability acting under given conditions."

of stimuli are brought together and given a somewhat unified effect in behavior."

(10) *Terman,* "Intelligence is the ability to think in terms of abstract ideas."

(11) *Thorndike,* "We may define intellect, in general, as the power of good responses from the point of view of truth or fact."

(12) *Woodrow,* "Intelligence is an inquiring capacity."

[3] Courtis, S. A., *Why Children Succeed.* Detroit : Courtis Standard Tests, 1925.

A critical consideration of these basic concepts which underlie all mental measurement makes it apparent that mental tests do not in any direct way measure the pupil's native capacity to learn, or even to any great extent his ability to learn. Rather, they measure what he has learned and when taken in relation to the individual's opportunity to learn and his rate of learning, afford most useful indirect evidence on capacity. Thus, the results of mental measurement reveal capacity indirectly, and are affected but not too much affected by the differentials of environment. Mental tests are incapable of securing a direct measure of capacity unaffected by experience and training, or the general native adaptability of the organism. Learning at any time is unquestionably affected to a large extent by experience. A pupil whether under self-direction or under teacher-direction, learns how to learn; and the results of his learnings, knowledges, and informations facilitate all subsequent learnings. It is quite obvious that differences in kinds of training and experiences affect the procedures of learning and that these differences will determine the fields in which learning and acquired abilities will be most efficacious. Mental tests do not test the actual process of learning, but test what has been learned through contact with environment.

141. **The Nature of the Mental Test.** For practical purposes general mental ability may be defined as the power of the individual to learn. This definition implies in a sense that this power to learn is a tangible, semi-objective mental quality which can be evaluated in a rather direct manner. Such is not the case, however, for power to learn can only be inferred or implied from the fact that learning *has taken place*. The value of the mental test lies in the fact that it affords an objective basis for this inference. It samples widely from the fields of learning resulting from experiences assumed to be common to all persons subjected to the test. The pupil's capacity to learn is determined by summing up his reactions to the items of the test. With a given opportunity for learning, the more tangible evidences there are that learning has taken place, the greater the implied power to learn. No intelligence test measures capacity directly, but within reasonable limits reflects the individual's power to

learn by measuring his ability to react to fragments of his environment. The quality of this reaction is evaluated generally in terms of the average reactions of large groups of unselected individuals. That is, most intelligence tests consist in batteries of different types of tests sampling into many different fields of interest and ability.

There is apparently no way of determining very precisely which particular fields of human interest or ability should be sampled in the attempt to secure this cross-section of mental activity. On this account there are very many different mental tests, and few of the authors of these tests agree with each other as to the proper content of such measuring instruments. It is important that the sampling be sufficiently diverse and representative to permit the securing of an estimate in the nature of an average which will not penalize a person because he may not have had this or that experience. Briefly, the measure or average obtained from a test which does sample representative reactions is taken to be truly indicative of one's ability to learn. Roughly it is assumed that what an individual has done with his mind by way of learning is indicative of the kind of mind he has. Differences in mental test scores are probably sufficiently accurate, rough as they are, to indicate such differences in mental power. Back of all of this is the generally accepted notion that the extent to which one has learned through experience is proportionate to his capacity to learn.

Teachers and educators in general are aware that there are at least two related but different phases of intelligence which must be taken into account in adequate mental measurement. Certain individuals react readily to abstract stimuli and thus are frequently rated as normal or even superior on the basis of certain mental tests in which abstract material predominates. Other types of individuals do not respond to abstractions but reveal unusual aptness in reacting to concrete and tangible material. Stenquist[4] has stated the case for this type of pupil most convincingly, and has presented a most useful supplement to the ordinary abstract type of mental measurement in his *Mechanical Aptitude Test.*

[4] Stenquist, John L., "The Case for the Low I.Q.," *Journal of Educational Research,* Vol. 4, pp. 241-254, November, 1921.

III. The Kinds of Mental Tests

Mental tests may conveniently be classified as individual or group tests. Individual tests are those administered to one person at a time; group tests may be given to a number of persons at the same time. Each type has its advantages and disadvantages. The use of one or the other type is dependent mainly upon the needs of the situation at the time of the testing.

142. **Individual Mental Examinations.** Individual mental examinations probably constitute the most accurate devices for the measurement of intelligence. The length of the test, the wide variety of reactions called for, the fact that the subject receives his instructions personally from the examiner, and the fact that this personal contact affords the examiner an opportunity to observe each reaction of the subject all combine to account for this greater accuracy of measurement. However, this greater accuracy is off-set by greater expense in the administraton of the tests. In fact, it is thought by many that this expense item is so great that in most classroom situations the resulting increase in accuracy of measurement is not commensurate with the added cost. Accordingly, it is quite probable that most classroom teachers will find it desirable to initiate their analysis of problem cases by first using good group tests of mental ability. Individual mental examinations may be given later to the relatively small number of cases who deviate most widely from the normal.

A very simple procedure will reveal directly to the teacher the special cases which should receive further attention. If the group mental test scores for the entire class are tabulated in a frequency table, or if the test papers themselves are merely arranged in descending order of size of the test scores, the individual pupils deviating most widely from the average for the class and from the normal mental age for the grade will be revealed. Thus, it may be necessary to retest or to give the individual mental examination to only a small percentage of the group. After all, it is in the highest and lowest extremes of intelligence that the problem cases arise, and it is also in this same group that the most serious errors

or mis-interpretations are likely to take place. Most present-day tests of mental ability accomplish a reasonably accurate placement of individuals in the more nearly normal group.

The individual test can be used to compare the intelligence of different persons, but only when they have had exactly the same learning opportunities, and when they are measured by the same instrument under exactly the same conditions. Tests capable of this type of discrimination are largely patterned upon the original Binet-Simon tests, an individual examination. Binet prepared the first edition of this mental examination in 1905 and revisions appeared rapidly after that. The 1908 and 1911 editions were prepared in collaboration with Dr. Simon. Decroly and Degand published a revision in Belgium in 1910, and the revisions by Goddard in America and Johnston in England appeared in 1911. Kuhlmann's revision was published in 1912. The most popular revision of this scale in this country, the Stanford Revision, was published by Terman in 1916. The Herring[5] revision appeared in 1922.

The general procedure in administering the Stanford Revision[6] is representative of that of many of the other revisions mentioned. The type of performance varies considerably with the different exercises. These exercises are presented to the child by means of spoken directions. The examiner tests one child at a time. The test should be given in a quiet room where there is freedom from distraction, and a friendly attitude between examiner and subject should be maintained. The examiner is directed to make sure that the subject understands what is to be done, and in all cases the burden of proof is with the examiner to show that the subject has responded in a way that is representative of his ability. The following brief quotation from Terman[7] emphasizes the importance of the examiner's getting into "rapport" with the subject to be examined. "The examiner's first task is to win the confidence of the child and overcome his timidity. Unless rapport has first been established, the results of the first tests given are likely to be misleading. The time and effort neces-

[5] World Book Company.

[6] Terman, L. M., *The Measurement of Intelligence,* Houghton Mifflin Company.

[7] Terman, L. M., *The Measurement of Intelligence,* Houghton Mifflin Company, 1916; page 124.

sary for accomplishing this are variable factors, depending upon the personality of both the examiner and the subject. In a majority of cases from three to five minutes should be sufficient, but in a few cases somewhat more time is necessary."

The chief advantage of the individual mental examination is the opportunity it affords for a trained examiner to observe the child's behavior under standard conditions. Since only one person can be examined at a time the individual mental examination is quite expensive in terms of time. Workers in this field, led by the work of Terman, Otis, Thorndike, Yerkes, and others, and stimulated by the psychological testing during the World War, finally developed very satisfactory instruments for measuring large groups of subjects at the same time. Of course, when group tests are given the difficulties and mistakes of individuals are not so readily observed by the examiner. In the individual examination the examiner may guide the subject into the best possible response for that subject. Along with this advantage there is the disadvantage that individual tests demand the services of highly trained examiners. Training in the field of psychology and mental testing is highly desirable from the standpoint of the administration of the tests and the interpretation of the results obtained.

The point to remember is that the correct interpretation of an individual's reactions under testing conditions is an expert matter and should be attempted only by a trained person. Not every teacher in a school need become a trained mental examiner. In fact it is probable that only a few are gifted with the patience, tact, and powers of observation necessary to enable them to become successful examiners. Many school administrators who cannot afford a full time psychological examiner are provided with means by which one or two carefully selected teachers can receive the proper training in this important skill and yield sufficiently expert service to the school system as a whole.

143. **Group Tests of Mental Ability.** Group tests for measuring intelligence include all those mental tests which may be given by an examiner to a number of children at the same time. The outstanding advantage of these measuring

instruments in comparison with those which must be administered to children individually is the great economy of time resulting from testing children in groups. There is an added advantage in that less training and experience are required of the examiner. On the other hand, there is the disadvantage of slightly less reliable results from group tests. A few of the more commonly used tests of this type are: the *Terman Group Test of Mental Ability*, the *Otis Group Intelligence Test*, the *Otis Self-Administering Tests*, the *Illinois Examination*, the *Haggerty Intelligence Examinations*, the *Kuhlmann-Anderson Intelligence Tests*, and the *National Intelligence Tests*.[8]

While the foregoing intelligence examinations have been listed together and no attempt has been made to classify them as to function, it must be borne in mind that their functions are not identical. Some emphasize performance while others sample more heavily into learned reactions. Some require careful administration in accordance with detailed directions while others are largely self-administering. Some are much more appropriate for use in the upper grades and high school than others.

Because they are representative of a large number of tests of this type, brief extracts from the *Kuhlmann-Anderson Intelligence Tests* and the *Otis Self-Administering Test of Mental Ability* are reproduced here with the permission of the publishers.

Kuhlmann-Anderson Intelligence Tests[9]

The battery of thirty-nine tests comprising the *Kuhlmann-Anderson Intelligence Tests* is arranged in a scale of overlapping units, the net effect of which is to approximate closely the results from any good individual mental examination. The tests in their most recently revised form are arranged in the booklets as shown in Table 41.

[8] More exact information concerning these tests and others named in this and in succeeding chapters may be obtained by studying the catalogues of the chief publishers and distributors of tests. A list of test publishers and distributors is given on pages 585-595 of the *Appendix*.

[9] Kuhlmann, F., and Anderson, Rose, *Kuhlmann-Anderson Intelligence Tests*. Minneapolis and Philadelphia: Educational Test Bureau, Inc., 1927.

TABLE 41

ARRANGEMENT OF KUHLMANN-ANDERSON TESTS

School Grade	Tests	Age Where Test Fits Best
I - 1 sem.	1 - 10	6 - 0
I - 2 sem.	4 - 13	6 - 6
II	8 - 17	7 - 6
III	12 - 21	8 - 6
IV	15 - 24	9 - 6
V	18 - 27	10 - 6
VI	22 - 31	11 - 6
VII - VIII	25 - 34	13 - 0
IX - Adult	30 - 39	15 - 6

The tests in this rather unusual examination are scaled according to the difficulty of response found at each successive age level. The booklet for the first semester of the first grade is made up of Tests 1 to 10 plus two practice exercises. Since more difficult items can be answered in the second semester of the first grade the booklet for this grade contains Tests 4 to 13 inclusive. In this way each successive grade booklet overlaps the one previous and insures a range of material for reaction suitable to the ability of the group. The booklets for Grades VII and VIII include Tests 25 to 34 inclusive. For the ninth grade and above, the booklet containing Tests 30 to 39 should be used.

A somewhat novel procedure is used in interpreting the results of these tests. Each of the ten tests comprising a booklet is standardized separately. The test is scored in terms of the number of exercises answered correctly. Referring to a table of norms shows the mental age of the individuals making a given score. A mental age score for each of the ten tests is thus obtained, the final mental age score assigned to the pupil being the median of the resulting mental ages. This procedure appears to result in unusually reliable measurement of mental ability.[10]

[10] Kuhlmann, F., "A Median Mental Age Method of Weighting and Scaling Mental Tests." *Journal of Applied Psychology*, June, 1927.

Test 35. Kuhlmann-Anderson Intelligence Tests
(For use in Grade Nine and above)

Write one letter on the line after each one of these statements:

Write A	If always true, as	Birds are hatched from eggs	A
Write B	If true in some cases, as ..	Snow melts as it falls	B
Write C	If not true, as	There are eleven inches in a foot.	C
Write D	If only an opinion, as	Skating is better sport than swimming	D

1. The lizard is a kind of insect 1
2. The largest cities in the world are seaports 2
3. A boy's uncle is the brother of one of his parents 3
4. A city is the best place in which to live 4
5. Very light boats sometimes float up-stream 5
6. Water boils more quickly in a high altitude 6
7. A concert is the best form of entertainment 7
8. Civil war is strife between different nations 8
9. People lose their self-control in a crisis 9
10. Capital punishment should be abolished 10
11. The largest city in a state is the capital 11
12. War greatly increases a nation's indebtedness ..12
13. The most valuable advice is disregarded 13
14. Anarchy is a means of promoting better government 14
15. It is easier to swim in salt water than in fresh water 15
16. Limitation of armaments would prevent war 16
17. More than half of the twelve months have 31 days 17
18. Holders of political offices are corrupt 18
19. Immigration should be restricted 19
20. One's temperature becomes lower in case of a fever 20

Test No. 35.

Directions for Test 35. Kuhlmann-Anderson Intelligence Test

Examples: "On this page you have the directions printed on the page. You are to write one letter on the dotted line after each one of these statements. Notice the example. Write A on the line if the statement is always true, as 'Birds are hatched from eggs.' Write B if the statement is true in some cases, as 'Snow melts as it falls.' Write C if the statement is not true, as 'There are eleven inches in a foot,' and write D if the statement is only an opinion, as 'Skating is better sport than swimming!'"

Trials: Allow two minutes. Then say, "Stop. Turn to next page. Pencils up."

"Begin with the first one and write the correct letter on the line after each one. Ready. Go."

Otis Self-Administering Test of Mental Ability[11]

This test is unique in that it requires only a minimum amount of direction from the examiner. For this reason it has proved to be a very satisfactory type of test for use in situations where pupils and teachers are not widely experienced in taking and administering tests.

Higher Examination. For High Schools and Colleges

This is a test to see how well you can think. It contains questions of different kinds. Here is a sample question already answered correctly. Notice how the question is answered :

Which one of the five words below tells what an apple is ?
1 flower, 2 tree, 3 vegetable, 4 fruit, 5 animal . . . . . . (*4*)

The right answer, of course, is "fruit" ; so the word "fruit" is underlined. And the word "fruit" is No. 4 ; so a figure 4 is placed in the parentheses at the end of the dotted line. This is the way you are to answer the questions.

If the answer to any question is a number or a letter, put the number or letter in the parentheses without underlining anything. Make all letters like printed capitals.

The test contains 75 questions. You are not expected to be able to answer all of them, but do the best you can. You will be allowed half an hour after the examiner tells you to begin. Try to get as many right as possible. Be careful not to go so fast that you make mistakes. Do not spend too much time on any one question. No questions about the test will be answered by the examiner after the test begins. Lay your pencil down.

Page 2

Examination begins here :

1. The opposite of hate is (?)
 1 enemy, 2 fear, 3 love, 4 friend, 5 joy . . . . ( )
2. If 3 pencils cost 5 cents, how many pencils can be bought for 50 cents ? . . . . ( )
3. A bird does not always have (?)
 1 wings, 2 eyes, 3 feet, 4 a nest, 5 a bill . . . ( )
4. The opposite of honor is (?)
 1 glory, 2 disgrace, 3 cowardice, 4 fear, 5 defeat ( )

[11] Otis, A. S., *Self-Administering Test of Mental Ability.* Yonkers, New York : World Book Company, 1922.

5. A fox most resembles a (?)
   1 wolf, 2 goat, 3 pig, 4 tiger, 5 cat . . . . ( )
71. A man who is influenced in making a decision by preconceived opinions is said to be (?)
   1 influential, 2 prejudiced, 3 hypocritical, 4 decisive, 5 impartial . . . . . ( )
72. A hotel serves a mixture of 2 parts cream and 3 parts milk. How many pints of cream will it take to make 15 pints of the mixture? . . ( )
73. What is related to blood as physics is to motion?
   1 temperature, 2 veins, 3 body, 4 physiology, 5 geography . . . . ( )
74. A statement the meaning of which is not definite is said to be (?)
   1 erroneous, 2 doubtful, 3 ambiguous, 4 distorted, 5 hypothetical . . . . ( )
75. If a wire 20 inches long is to be cut so that one piece is $\frac{2}{3}$ as long as the other piece, how many inches long must the shorter piece be? ( )

IV. Educational Uses of Mental Tests

144. **For Class Diagnosis.** Viewed from the standpoint of the classroom teacher, achievement tests and intelligence tests supplement each other. After the teacher has given achievement tests and compared his class with the norms in a given subject he is still in danger of making several false assumptions relative to the significance of these results unless he has available further information such as is furnished by intelligence tests. He may credit himself with an excellent job of teaching, when the innate brilliance of the group in his charge is such that if given really adequate instruction much superior results might have been achieved. It may be that the class falls considerably below the expected norm, and the teacher may feel that his teaching has proven a failure. This may also be an unwarranted assumption, for the class may be considerably below average in intelligence and can not be expected to approximate the norms that are set up for a class of average ability. It is evident then that there is a need for some means of determining approximately the intellectual capacity of the class. Mental tests meet this need. By giving one or more such tests, the teacher can determine with a fair degree of accuracy whether his classes are up to

the normal expectation in ability to master school work. The teacher is almost certain to encounter a greater number of technical difficulties in both giving and interpreting mental tests than he finds in using achievement tests. However, with practice and with the use of proper precautions he may come to have confidence that his conclusions are approximately correct for the class as a whole. He may, by comparing the results of several of these group mental tests, make a diagnosis of the ability of his class which will go a long way toward enabling him to determine what degree of proficiency he should reasonably expect in connection with the teaching of the regular subject matter. This knowledge will enable him to plan his work more intelligently, and to lay plans for overcoming difficulties and barriers which might otherwise have been unsuspected or delayed in making their appearance. Usually intelligence testing is done by the supervisory force, and the teacher simply calls upon the proper authorities for this item of information. While achievement testing may be thought of as essentially a part of the work of the classroom teacher, mental testing is more likely to be considered the work of the specialist, especially in the case of the individual mental examination.

145. **For Individual Diagnosis.** The mental test may prove especially valuable to the classroom teacher in assisting him to solve the problems relating to the unusual child. The pupil may be unusually bright, troublesome, dull, mischievous, or in some other way quite out of the ordinary. The teacher may wish to know whether this child's typical responses actually reveal his real general ability, and whether or not the judgments of his former teachers and supervisors are correct. Mental tests will give information not obtainable in any other way. The results may indicate that the child has ability hitherto unsuspected; or that his supposed brightness is but a superficial pertness covering a really dull intellect. If it is a case of misconduct it may show that the child's bad behavior comes from the failure of the teacher to keep busy a really brilliant intellect, thus allowing his superfluous mental activity to find its outlets in mischief and disorder.

The mental test, when given to an entire group, frequently

uncovers a child of outstanding capacity who has been content to go on with the group without especially revealing his real ability. Many such cases would never have been revealed, had it not been for the mental test. Mental tests reveal cases of overlapping in ability just as achievement tests reveal cases of overlapping in school achievement. They may reveal children in the seventh or eighth grades with the intelligence of normal twelfth grade pupils, and may also reveal eleventh or twelfth grade pupils who in intelligence should be in the seventh or eighth grade.

When children are discovered who are far in advance of their place in school, readjustments of work should be made to match their abilities. This may be accomplished in some cases by advancing them to a grade of work where their intelligence is given a real test, or by placing them in rapidly moving classes, so that they may progress according to ability rather than by some fixed promotion scheme; or by declaring for the entire class minimal requirements in the way of the different units of work to be done and then expecting the brighter pupils to attack the various problems at deeper levels and more intensively than could be expected of the class as a group.

By proceeding along similar lines the dull pupil will also be better understood, because his difficulties can then be diagnosed more particularly and his strong points brought into relief, resulting quite frequently in an entire re-direction and reorganization of his instruction. In any case, the mental test will assist both in explaining difficult cases, and in revealing unsuspected general strengths and weaknesses.

146. **Care in Use of Mental Test Results.** On the whole, intelligence tests seem secure in the place they now hold as indispensable supporting agencies for achievement tests, as valuable instruments for the more exact classification and gradation of pupils in the classroom, as real guides to the teacher in matters of behavior and conduct as well as to the child himself in certain matters of vocation. There are, however, a few dangers attached to their indiscriminate use which the careful teacher and administrator should guard against. The more important of these are probably social in their character. In the first place, there is the danger

which may arise through giving publicity to the results of intelligence testing. There is probably no other place in the use of educational measurements where so much real "social dynamite" is to be found. The parents of children are reasonably tolerant of mental testing devices until their own children are measured. If their children are tested in arithmetic and show up poorly, it is a matter of small consequence for that is the teacher's fault. It is a mere matter of faulty instruction and reflects on the school alone. But let one of their children show up poorly on a mental test and it is a different matter. Capacity is inherited and the results then do reflect on the parents. In the long run nothing but damage will be done by using intelligence test results for any other than purely administrative purposes and these in the strictest confidence.

A second danger in the careless use of intelligence tests lies in the effect which knowledge of such matters may have on the individual himself. The fact that a child is publicly labelled as unusually brilliant may release a flood of egotism on his part, or, depending upon his personality, subject him to the gibes of his fellow students. The fact that a child is publicly labelled as slow mentally may place a stigma on him which can never be removed. The only safe practice is to refuse consistently to divulge the results of such examinations to anyone not in a responsible administrative position and then to insist strictly upon their confidential use. Absolute adherence to this rule is the only safeguard, because any exception will set a precedent. The value of the tests to the child and to the teacher is seriously reduced by an injudicious use of the results.

V. Inter-dependence of Mental and Achievement Tests

147. Inter-dependence of Achievement and Mental Test Results. The teacher should obviously not attach too much importance to the results of intelligence tests or achievement tests alone. *Mental tests should supplement and not supersede the results of measurement of classroom achievement.* They should not be used in place of all other obtainable in-

formation about a pupil or a class, but should be interpreted in the light of such information. Accordingly, in order that too great stress may not be laid upon the child's proficiency in any one sort of test, the well informed supervisor or teacher attempts to rate the pupil in terms of his various attainments. By the intelligence test, the teacher discovers his mental "age," that is, how his general ability to learn compares with the norm for normal children of any given chronological age. The teacher may find that a child 10 years and 6 months old, actually tests on the mental test as high as the "normal" child of 12 years and 3 months. This child is then said to have a "mental age" of 12 years and 3 months. His *I. Q.* (Intelligence Quotient) is the ratio of his mental age to his chronological age. In this case it is the ratio of 12 years 3 months to 10 years 6 months. Reducing to months, the ratio becomes 147 to 126 or 1.16. In actual usage it is the practice to multiply this figure by 100, expressing the *I. Q.* as the whole number 116. This means that this child is considerably accelerated in his mental development.

But let us assume the teacher has also tested the child in other ways. His ability has been measured on a battery of achievement tests such as the *Unit Scales of Attainment* or on the *New Stanford Achievement Tests*. The teacher finds the pupil's achievement age in these subjects, as compared with his chronological age, and works out ratios for these subjects, as between his chronological (real) and mental ages.[12] The norms for most modern achievement tests are given in terms of years and months.

Thus the achievement age of the child may then be compared with his real or life age. The resulting quotient is called the *E. Q.* or the "Educational Quotient." If the *E. Q.* is then divided by the *I. Q.* of the pupil, the result is called

[12] McCall in his "How to Measure in Education" presents a table based upon the investigations by Ayres, Terman, and Kelly which gives the average age of a first grade child as 80 months, at the entrance into school, and adds 13 months to this figure for each succeeding year, as the investigators quoted seemed to agree that the average time spent in each grade is from 12–13 months. On this basis the average age of a sixth grade pupil is 145 months, or if the norm is for the month of May instead of September, as will often happen, then the age will be 154 months.

the *A. Q.* or the "Achievement Quotient."[13] Franzen[14] has suggested that the accomplishment quotient be calculated for a combination of achievement tests instead of being based upon a single test. Buckingham and Monroe[15] have devised a set of combined educational-mental tests arranged so as to facilitate the calculation of the *A. Q.* Tests of this type give, of course, rather rough measures, but have been used rather extensively for solving the problems of group diagnosis.

VI. Derived Results of Mental Testing

Derived measures based on mental test results are fundamental in any attempt to reveal the interdependence of achievement and mental ability.

148. **Mental Age As a Derived Measure.** Terman defines mental age as "that degree of general ability which is possessed by the average child of corresponding chronological age." It is "an index of absolute mental level" and "indicates the level of development which a child has reached at a given time."[16] For example, a child has a mental age of ten years when his level of mental development (mental ability) is equal to that of the normal child of exactly ten years. Thus, if a representative group of pupils, all of whom are ten years of age, take an intelligence test and earn an average point score of 45, any pupil who subsequently takes this test and earns a point score of 45 is said to have a mental age (M.A.) of ten years. In the same manner, for each age group an average point score is established. By comparing point scores earned by pupils with a table of mental-age norms, the mental age of each pupil can be determined.

149. **Intelligence Quotients as Derived Measures.** When the chronological age (life age in years) of a pupil is known

[13] See discussion on pages 241-243 of Chapter IX.

[14] Franzen, R., "The Accomplishment Quotient," *Teachers College Record*, November, 1920.

[15] *Illinois Examinations*, Public School Publishing Company, Bloomington, Illinois.

[16] Terman, Lewis M., *The Intelligence of School Children*, Chapter I. Boston: Houghton Mifflin Company, 1916.

and his mental age has been determined from his point score on a mental test, his intelligence quotient, or I.Q., may be computed. The intelligence quotient is a simple method of expressing the relationship between a pupil's mental age (M. A.) and his chronological age (C. A.).

The procedure for computing the I.Q. is as follows: the total number of months in the child's mental age is divided by the total number of months in his chronological age. The result is the intelligence quotient. For a pupil with a C. A. of ten years and five months, with an M. A. of twelve years and six months as shown by the test, the computation of the I.Q. would be as follows:

$$\frac{\text{M. A. in months (150)}}{\text{C. A. in months (125)}} = 1.20 \text{ (I.Q.)}$$

The intelligence quotient is expressed as 120 — the decimal point is omitted and the I.Q. is considered as a percentage. In this case the pupil's mental age is 120% of his chronological age.

It should be kept in mind that the mental age indicates the mental level which the pupil has attained at the time of the test, while the I. Q. is an index of his relative brightness. Mental age signifies the pupil's power at the time he took the test, while the I. Q. signifies his relative mental development for his age. To the degree the I. Q. remains constant, it signifies future possibilities with reference to mental accomplishment

150. The Personal Constant (P. C.) as a Derived Measure. For some time psychologists have expressed dissatisfaction concerning the constancy of the I. Q. The best evidence indicates that in general the I. Q. does not remain constant throughout successive years, except for children with an initial I. Q. of approximately 100. The records show that there is a consistent tendency for initial intelligence quotients below to decrease with increase in age, and for initial intelligence quotients above 100 to increase with increase in age. Kuhlmann writes: "The amount of change per year depends on the initial I. Q. and on the chronological age. The total rate during the mental growth period may exceed 25 I. Q. points. The change is due to the fact that the rate of

normal or average mental development slows up with increasing years."[17]

Dr. H. Heinis of the University of Geneva became much interested in this problem and secured data on the mental development of children ranging in age from two to eighteen years.[18] By combining his material with that of other investigators he finally derived a formula for a curve of nor-

TABLE 42

HEINIS MENTAL GROWTH UNITS[19]

YEARS	MONTHS											
	0	1	2	3	4	5	6	7	8	9	10	11
0	0	5	10	16	21	26	31	36	41	46	51	55
1	60	64	68	73	78	82	86	91	95	99	103	107
2	111	115	119	123	127	130	134	138	141	145	148	152
3	155	159	162	165	169	172	175	178	181	184	187	190
4	193	196	199	202	205	208	210	213	216	218	221	224
5	226	229	231	234	236	238	241	243	245	248	250	252
6	254	257	259	261	263	265	267	269	271	273	275	277
7	279	281	282	284	286	288	290	291	293	295	296	298
8	300	301	303	304	306	307	309	310	312	313	315	316
9	318	319	320	322	323	324	326	327	328	329	331	332
10	333	334	335	337	338	339	340	341	342	343	344	345
11	346	347	348	349	350	351	352	353	354	355	356	357
12	358	359	360	361	361	362	363	364	365	365	366	367
13	368	369	369	370	371	372	372	373	374	374	375	376
14	376	377	378	378	379	379	380	381	381	382	383	383
15	384	384	385	385	386	386	387	387	388	388	389	389
16	390	390	391	391	392	392	393	393	394	394	395	395
17	395	396	396	397	397	397	398	398	399	399	399	400
18	400	400	401	401	401	402	402	402	403	403	403	404
19	404	404	405	405	405	406	406	406	406	407	407	407
20	408	408	408	408	409	409	409	409	410	410	410	410
21	411	411	411	411	411	412	412	412	412	413	413	413
22	413	413	414	414	414	414	414	414	415	415	415	415
23	415	416	416	416	416	416	416	416	417	417	417	417
24	417	417	418	418	418	418	418	418	418	418	419	419
25	419	419	419	419	419	419	420	420	420	420	420	420
26	420	420	420	421	421	421	421	421	421	421	421	421
27	422	422	422	422	422	422	422	422	422	422	422	423

28-0 to 29-0	423	32-2 to 34-3	426	Over 45-0	429
29-1 to 30-4	424	34-4 to 37-8	427		
30-5 to 32-1	425	37-9 to 45-0	428		

[17] *Instruction Manual for Kuhlmann-Anderson Intelligence Tests.* Minneapolis and Philadelphia : The Educational Test Bureau, Inc., 1933.

[18] Heinis, H., "La loi du Developpement Mental." *Archives de Psychology,* Vol. 74, 1924.

[19] Computed from Heinis' formula by A. H. Hilden. Reproduced with the permission of the author and the publishers of the Kuhlmann-Anderson Intelligence Test Instructional Manual. Educational Test Bureau, Inc., Minneapolis and Philadelphia, 1933.

mal mental growth and for a coefficient which he believed to be superior to the I. Q. as an expression of the relative development of intelligence.

The Heinis formula[20] is intended to give quantitative expression to the normal mental growth curve in terms of growth units for each age level. A table of the growth units for each age level is shown on page 277. These units were derived by means of the Heinis formula.

To compute the Heinis Personal Constant the mental age is first required, as for the I. Q. It is secured in the same manner and from the same intelligence tests. The Personal Constant (P. C.) of a pupil whose age (C. A.) is 10 years 3 months and whose mental age (M. A.) is 12 years 9 months would be found as follows: By using Table 42 the M. A. of the pupil is turned into growth units. A mental age of 12 years 9 months is equivalent to 365 growth units. This value becomes the numerator of the fraction. According to Table 42 chronological age of 10 years 3 months is equivalent to 337 growth units. This value becomes the denominator of the ratio on which the P. C. is based. By dividing 365 by 337 and pointing off two places to the right as is always done in quotients in order to express them as whole numbers the P. C. is found to be 108. The I. Q. $\left(\frac{\text{M.A.}}{\text{C.A.}}\right)$ of this pupil with a chronological age of 10 years, 3 months (123 months) and a mental age of 12 years, 9 months (153 months) would be 124.

Comparisons of intelligence quotients and personal constants have shown them to be strikingly different in range and in character for many age levels or school grades. This is true for either dull or bright pupils, but especially so for those of high intelligence. Kuhlmann is so convinced of the superiority of the Heinis Constant that he strongly recommends its use in place of the I. Q. and now provides a table of Heinis Mental Growth Units for use in connection with the *Kuhlmann-Anderson Intelligence Tests.*

[20] Heinis, H., "A Personal Constant." *Journal of Educational Psychology*, Vol. XVII, pp. 163-186, March, 1926.

VII. Summary

The nature and the meaning of mental ability as related to the problems of instruction are discussed in this chapter. Intelligence is defined, and in the light of this definition the meaning of the mental test score is stated. The major types of mental tests are considered from the standpoint of their comparative, instructional, and guidance functions. A few of the more valuable mental tests for classroom use are discussed in detail. Certain danger points in their use are definitely emphasized. Some of the newer techniques in the interpretation of mental test results are also presented. The reader who is interested in making a more critical study of mental testing will find much help in the references at the end of this chapter.

Discussion Exercises

1. From your reading on the subject state a definition of intelligence which is acceptable to you as a student.
2. Show how a high performance score on a mental test affords a basis for the implied existence of capacity of a correspondingly high level.
3. Discuss the truth of the statement: "An individual mental examination gives a complete picture of the total mind of the subject."
4. Make a list of specific advantages and disadvantages of individual intelligence examinations in comparison with group mental tests.
5. Point out a few of the limitations of group tests of mental ability when used in the high school grades.
6. What is the difference between the Mental Age (M.A.) and the Intelligence Quotient (I.Q.)?
7. What do you understand to be the significant differences between the Intelligence Quotient (I. Q.) and the Personal Constant (P. C.)?
8. Show how the supplementary use of intelligence test results protects the teacher against the misinterpretation of achievement test results. How is the Accomplishment Quotient related here?
9. After looking over samples of group mental tests make a list of at least ten types of test exercises which in your judgment

would be suitable for use in a group mental test for a specified grade.

10. Pupils A and B in the sixth grade have been doing barely passing work in arithmetic. Suppose that each pupil was given an individual mental examination with these results :

Pupil	Chronological Age	Mental Age	*I. Q.*
A	12 yr. 6 mo.	15 yr. 10 mo.	?
B	13 yr. 6 mo.	10 yr. 1 mo.	?

Assume that the parents of these two pupils call upon you regarding their school work. Write out a statement of what you would tell each parent. Would you state the I. Q. of these children ? In what grades do these children properly belong as indicated by their mental test scores (assume 12 years and 6 months as normal for the sixth grade) ?

Selected References

Cattell, Psyche, "The Heinis Personal Constant as a Substitute for the I. Q." *The Journal of Educational Psychology*, Vol. 24 : 221-228, March, 1933.

Dearborn, W. F., *Intelligence Tests, Their Significance for School and Society.* Boston : Houghton Mifflin Company, 1928.

Freeman, F. N., *Mental Tests, Their History, Principles and Applications.* Boston : Houghton Mifflin Company, 1926.

Heinis, H., "A Personal Constant." *Journal of Educational Psychology*, Vol. 17 : 163-186, March, 1926.

Herring, J. P., *Herring Revision of the Binet-Simon Tests.* Yonkers, New York : World Book Company, 1922.

Kuhlmann, F., Anderson, Rose, *Instruction Manual for the Kuhlmann-Anderson Intelligence Tests.* Minneapolis and Philadelphia : The Educational Test Bureau, 1933.

Pintner, R., *Intelligence Testing.* New York : Henry Holt and Company, 1931.

Riley, Gordon, "A Comparison of the Personal Constant and I. Q." *Psychological Clinic*, Vol. 18 : 261-265, February, 1930.

Spearman, C., *The Abilities of Man.* New York : The Macmillan Company, 1927.

Symposium, "Intelligence and Its Measurement." *Journal of Educational Psychology,* Vol. 12 : 271-275, 1921.

Terman, L. M., *Group Test of Mental Ability, Examiner's Manual.* Yonkers, New York : World Book Company, 1920.

Terman, L. M., and others, *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence.* Baltimore, Maryland : Warwick and York, 1917.

Terman, L. M., *The Intelligence of School Children.* Boston : Houghton Mifflin Company, 1919.

Thorndike, E. L., and others, *The Measurement of Intelligence.* New York : Bureau of Publications, Columbia University, 1927.

CHAPTER XII

MEASUREMENTS AND THE TOTAL ADJUSTMENT OF THE PUPIL

THIS chapter brings together the following points which appear to have a bearing on the total adjustment of the individual pupil:

a. Appraisals of total personality basic to pupil adjustment.
b. School responsibility for pupil adjustment.
c. Functions of measurement for adjustment and guidance.
d. The adjusted pupil.
e. Measurement of personality and character traits.
f. Measurement of vocational interests.
g. Measurement of specific aptitudes.

I. EDUCATION AS ADJUSTMENT

151. **Adjustment and Guidance.** Education is concerned with the problems of aiding the individual in making the best possible adjustments to his environment, present and future. Thus, practically, education *is* guidance and adjustment. Unfortunately, many institutions and teachers are so concerned with the teaching of subject matter that they forget their responsibilities for the development of personality. Furthermore, many of the adjustments required of these developing personalities are so complicated that teachers frequently find themselves ill-equipped to direct them. The teacher, or guide, presumably is one who has been over the route before, and thus knows something of the experiences and the difficulties to be encountered. However, personality and character traits are so many-sided that it is too much to expect an individual teacher to be able to give the pupil equally good guidance in making adjustments in all phases of his environment. Even the most richly informed teacher must have supplementary informa-

tion if this guidance is to be even reasonably effective. It is at this point that certain educational measuring devices make many helpful contributions to the teacher and to the individual pupil himself.

152. **School Responsibility for Adjustment.** The school through its administrative officers and its teachers faces a very real responsibility for the proper guidance and adjustment of its pupils. Correct guidance, if generally practiced, would pay big dividends in a more effective society and in happier and in better adjusted individuals. Not everyone requires assistance in making the required adjustments to life, but those who do usually need it badly. All agencies engaged in rendering social service face a definite responsibility for the development of a sane and comprehensive guidance program. The common types of guidance are vocational, educational, mental, physical, civic, moral, avocational or cultural, and social. While these forms of guidance are frequently administered by different agencies, they are highly related and involve many identical problems. For the school and the teacher these problems are particularly vital.

Exploratory courses frequently offered at the junior high school level open many possibilities for vocational adjustments. Pupil failure in many subject-matter fields, and the wasted time of pupils and the wasted efforts of the already overloaded teacher all argue strongly for the need for more and better educational guidance. If a pupil decides to become a doctor, a teacher, an engineer, or a lawyer, he needs guidance in the selection of the correct background courses as well as in the collection of information about the professions, the types of training required, the type of personality likely to succeed in the profession, and the possibilities of future rewards in these professions. Pupils who have made no definite selection of their life activity particularly need this type of information as a basis for a better appreciation of the requirements and the possibilities of the professions.

Increased mechanical efficiency in industry, shortened hours of labor, and frequently extensive unemployment have combined to increase the proportion of leisure time for each individual to a point where recreational guidance is a neces-

sity. In this field the assistance of industrial arts teachers and physical education teachers should be particularly helpful. There is great need for developing a clear understanding of the strains on the mental and physical fibres of youth due to the stress and speed of modern life. The guidance of a sympathetic doctor, school nurse, or if necessary psychiatrist should smooth out many of these difficulties. The social studies teachers should make their contribution to civic guidance. Moral guidance is a responsibility of the entire school and community. All such educational adjustment affords the teachers an excellent opportunity to modify character and develop desirable personality traits. This, of course, assumes that they know what desirable personality traits are, and are willing to accept their responsibility as teachers to take an active part in this most important aspect of teaching.

153. **Adjustment as Related to Personality, Character, and Individuality.** Personality and character are terms in such common use that it is difficult to believe that they lack clear definition. We speak glibly of both, yet we define them only vaguely. No other single outcome of our educational or social program is so important as the development of these traits. Every teacher, parent, or even pupil realizes the importance of good character, and a pleasing personality. But what is character? What is personality? Are they inherent or are they acquired? Can the teacher, parent or pupil do anything to further their proper development? The answers to these questions constitute real problems in the life philosophy of every teacher and student of education. The contribution which educational measurements makes is so great that no discussion of measurement in education would be complete without considering its bearing on these problems of the total adjustment of the child.

Personality may be considered as revealing itself through four major groups of traits; namely, physical, intellectual, temperamental, and character. The physical traits identified with personality involve an almost infinite number of possible combinations of such qualities as health, height, weight, energy, beauty, muscular tone and adaptability. The intellectual traits include general mental acuity, power to learn,

ability to evaluate and adjust oneself to new relationships. The temperamental qualities are much more vague. They may be regarded as the window through which the individual looks out upon the world and through which the world views the individual. The reaction of the individual to the world is colored by the nature of this window. In a similar way the world's evaluation of the individual is conditioned by his temperamental qualities. The fearful, the cringing, the fawning, the deceitful, the suspicious, the selfish, the pugnacious, the surly individual each casts his own peculiar shadow on the world and the world reacts to it.

The character traits are the instruments which steer the individual in his contacts with the world. Character has been defined as "the sum total of his (an individual's) behaviour in relation to the world about him and to his fellow beings."[1] Thus character may be taken as being almost synonymous with the behavior aspects of personality. Character is revealed by an individual's responses to stimulation. If the responses are acceptable or pleasing to his fellows the individual is reported as having a "good" character. If the reverse, he is regarded as having a "weak" character. Yet it is certainly true that a person's real character is revealed only when he is free to act without the inhibitions resulting from social pressure. Personality is revealed in the effect his responses have on his fellow beings. To a degree it is unfortunate that common usage has tended to attach to personality the notion of distinctiveness or individuality in character, for in their efforts to develop a distinct individuality many persons reveal character traits which are far from pleasing. Individuality is undoubtedly an important attribute of personality, but it is not so important that it should be achieved at the expense of honor, honesty, morality, bravery, modesty, or many other qualities which, in the past at least, have been considered desirable in an individual.

[1] Benson, C. E., Lough, J. E., Skinner, C. E., West, P. V., *Psychology for Teachers*. Boston : Ginn and Company, 1926.

II. Guidance As Adjustment

154. **Guidance, Vocational and Educational.** The general objective of guidance is to aid pupils in more wisely purposing, planning, executing, and evaluating all of the varied activities and interests which engage their attention. The aim of all guidance is to assist individuals in becoming happy and efficient human beings. Guidance is more than leading, conducting, regulating or directing persons, things or activities. These terms imply too much external and too little internal responsibility for the progress of the individual's own education. Accordingly, guidance must provide personal help that is offered in such a manner as to create in each individual the desire and the ability to carry on by his own efforts. The accomplishment of these objectives of guidance is not an easy task.

Guidance, according to the National Vocational Guidance Association, is "the process of assisting the individual to choose an occupation, prepare for it, enter upon it, and progress in it. As preparation for an occupation involves choice of studies, curricula, schools, and colleges, it becomes evident that educational and vocational guidance cannot be separated." The selection and preparation for an occupation is only one phase of guidance. On this point Jones[2] makes the following comment: "The very attempt to help young people in vocational choices inevitably reveals other needs that in many cases transcend in importance the vocational needs. Many times the boy is not ready to take a position; he needs further training. Often the need is a health need, a moral need, or a recreational need. Gradually but surely we have come to realize that guidance is not something which concerns only a part of the individual nor does it deal merely with a part of his life. The need is for a 'whole-child' guidance. At one time one particular need may be of greatest importance, at another time another may be paramount." Thus this wider view has gradually replaced the old idea of guidance. Just as life is much more than a

[2] Jones, Arthur J., *Principles of Guidance*. McGraw-Hill Company, Inc., New York, 1930. Pp. 367-368.

vocation, so guidance is and must be much more than vocational guidance.

The recognition that no two persons are alike is basic to effective guidance. The realization that different subject-matter fields and different occupational activities each make their specific demands on the individual is equally basic. Each individual's special aptitudes, abilities, interests, and personality qualities must be determined in order to assist him in the choice of a congenial vocation or field of study. This necessitates the analysis of fields of endeavor to determine just what demands each makes on its workers. The study of the individual includes collecting data on such items as:

1. The quality and quantity of his educational achievements.
2. The type of course work which claims his interest.
3. His physical qualifications and limitations.
4. His capacity to profit from training.
5. The quality of his character, and personality traits.
6. The social, educational, financial, and occupational background of his family.
7. His extra-curricular and recreational interests.
8. His chronological age and place in the family.
9. His specific vocational interests.

The analysis of the demands on the individual made by certain occupations and fields of study includes the following:

1. The determination of the nature of the work to be done.
2. The demands of the activity on the physical ability of the individual.
3. The demands of the activity on the intelligence, capacity, ability, temperament and character of the individual.
4. The opportunity it offers for advancement and training.
5. The immediate and ultimate possibilities of financial reward.
6. The conditions of labor.
7. The importance of the occupation or the activity in the field to which it belongs.
8. The number of persons engaged in the occupation.
9. The probability of the increase or decrease in the number of persons engaged in the occupation.

Each case constitutes a complex adjustment problem which must be treated as an entity. Individuals differ greatly in

ability. They differ in their power to learn and in their ability to remember. They vary markedly in specific aptitudes. They show wide differences in moral standards. They differ in physical size, in weight, in health, speed and strength of reactions, and in many other less tangible personality qualities. In addition to the normal variations extreme cases of maladjustment are sometimes encountered. These represent the most difficult problems of all. Pupil adjustment to the stimuli of our modern complex world requires the best and most objective information which can be obtained about the individual's inheritance and physical environment. To be effective, the individual must be reasonably attuned to the world about him. But effective living does not mean mere adjustment to a static or even to a changed environment. It means adjustment to an increasing range of environmental variations. When the children who are now in school reach maturity the world will be vastly different from the one they knew as children. So it has always been as the younger generations have grown to maturity. It is probable that today the movement is somewhat faster, but it is not otherwise very different.

Adjustment of the new member to society is a problem for the entire school and community. The job is far too big for any one teacher or individual to handle alone. It is an enterprise calling for the wisest co-operation of parents, teachers, and all of the constructive social agencies of the community. The intelligent adjustment of individuals to society is the most important educational obligation of a democracy to its citizens.

155. **The Adjusted Pupil.** The properly adjusted individual is one who has learned to fit himself smoothly and without warping his personality into the social group of which he would normally form a part. Proper adjustment of the individual implies that the different human qualities are properly balanced and combined in right relations into an integrated whole—the total personality. This total personality involves the happy integration of the physical, intellectual, moral, social, aesthetic and religious natures of each individual. This point of view treats normality as the criterion of adjustment. This in no sense means that the

teacher is to disregard the importance of individual differences or deviation from the norm. As a matter of fact it places a greater premium on their recognition.

Individuals vary slightly in almost every quality, and considerably in many. Normalcy depends upon the range of this variation. Tom, Dick, and Harry may vary in height, weight, intelligence, educational achievement, athletic prowess, sociability, temperament, in their interests, their pleasures, their ambitions, and in their general outlook on life. It is possible that when examined the three boys would be found to be quite normal and similar on one or more of these qualities but quite unlike in their total personalities. Probably the most desired condition for normalcy is to be born normal. After that, the best indication of normality in the individual is his power to keep growing physically, mentally, morally, and socially according to the normal rate. No one is completely normal in all respects, and probably would not be as valuable a personality, and surely not as interesting if he were.

From the standpoint of the school the adjusted pupil is the individual who is making normal and regular educational progress as compared with other individuals of similar capacities and traits. He is associated with groups of comparable mental level. He is adapted morally and socially to the individuals he meets in his daily routine. He is physiologically adjusted in that his sight, hearing, speech, and general health are comparable to the groups of which he is a part. The purpose of defining adjustment in this way is to make it apparent that the important thing is the process of fitting the individual into surroundings which are reasonably normal for him, rather than merely finding his absolute level of abilities or deficiencies. The mentally handicapped twelve-year-old child would be badly adjusted if placed with normal twelve-year-olds. The child with a speech defect may be maladjusted if placed with children with normal speech, but may readily adjust himself so far as his personality is concerned if he is placed with individuals by whom he is not submerged. This is a concept of prime importance to the teacher. The use of all available techniques and instruments for the collection and interpretation

of guidance information which will make possible this adjustment is an obligation which each teacher must assume.

156. **Basic Approaches to Pupil Adjustment.** As the basis for effecting the changes in the educational program necessary to bring about better pupil adjustment in the school, many types of accurate and detailed information concerning the individual are needed. In this connection a clear distinction should be made between facts and opinions. Although subjective estimates of a pupil's achievement during his earlier educational history may be of significant value in a guidance program, they are far less useful than objective scores from valid and reliable measuring instruments would be, other things being equal. Prominent among the basic approaches to pupil adjustment, therefore, is the matter of the availability of sufficiently accurate and extensive information about the individual. Of almost equally great importance is the need for a sympathetic and tolerant interpretation of this guidance information.

157. **Sources of Guidance Data.** Among the more important sources of information for guidance purposes is the complete record of the individual's educational history, including the studies pursued, achievement in these subjects, participation in extracurricular activities, acquisition of special honors and awards, and regularity of school attendance. Racial characteristics, the home life, and the social and economic conditions in which the pupil lives are also worthy of investigation. Data on health and physical conditions are frequently available from school records. Scores on intelligence tests, specific aptitude tests, and ratings on vocational interest and personality blanks can usually be obtained. In extreme cases of mental, moral or social maladjustments a psychiatrist may be consulted.

Cumulative Records of Educational History. Most modern school systems now require some form of permanent record card bearing facts and judgments of value for either immediate or later use. Such records, if truly cumulative and covering a reasonable section of the individual's educational history, provide most valuable guidance data. The importance of the school record is clearly stated in the following extract from Bulletin Number 19 of the *National*

Association of Secondary School Principals:[3] "The purpose of records is to reveal the conditions that exist in the school regarding the pupil, and to furnish such information as the school and the home may need in the full development of the pupil's educational possibilities. The information must of necessity relate to the intellectual, physical, moral, social, and spiritual growth of the individual pupil."

The committee on guidance appointed by the *National Association of Secondary School Principals* has recommended a complete series of record forms which should provide significant guidance information. The following forms are recommended:

a. Cumulative Scholarship Record
   Scholastic marks by subjects, semesters, teachers, etc.
b. Conference Data and Records
   Summaries of school records, estimates by classroom teachers, records of conferences with student and parents. Of special use to the guidance counselor.
c. Personal Record
   Summary of personal traits, social, and extra-curricular activities and interests, and vocational experiences.
d. Record of Tests
   Summary of data from intelligence, achievement and prognostic tests.
e. Adjustment Record
   Course and curriculum changes.
f. Placement Record
   Follow-up record of recommendation, employment, etc.
g. Health and Physical Record
   Family history, diseases, vaccinations.

The student who is especially interested in this aspect of pupil guidance will find a very helpful discussion in Chapter XIII of *Principles of Guidance* by Professor Arthur J. Jones.[4] Cumulative record forms have been proposed in connection with Child Accounting programs in a number of states. The space requirements of such forms prevent their presentation here.

Sympathetic Interpretation of Guidance Data. The collection of data pertaining to the individual's educational

[3] National Association of Secondary School Principals, Committee on Guidance, Guidance in Secondary Schools. Bulletin No. 19, Cicero, Illinois, 1928.

[4] Jones, Arthur J., *Principles of Guidance.* McGraw-Hill Book Company, New York, 1930.

background and progress is a useful step, but taken alone it is insufficient. Intelligent interpretation of the data must follow. Sympathetic guidance must take into account the fact that frequently there are serious conflicts between the individual's needs and desires and the group demands to which he is expected to respond. Guidance must not aim to make all equal or alike. Its aim is to aid the individual in finding himself, and in becoming an effective participant in society at large. This means the adjustment of the entire individual, in all of his interests, capacities, and abilities. It should strive continually to assist the pupil in his efforts to grow, to develop, to find, to discover, and to create himself. Persons engaged in this type of educational service must have a sympathetic and tolerant outlook on life, and must be trained in the application of this philosophy in handling their guidance problems.

Objective Evidence for Guidance Purposes. Certain groups of qualities contributing to personality lend themselves reasonably well to objective measurement. Others are extremely difficult to identify. Such physical traits as height, weight, available energy, physical aptitude or dexterity have been rather definitely catalogued by workers in the field of physical education. Phases of these traits which are measurable with reasonable objectivity are discussed in Chapter XX.

The meaning of intelligence and the possibilities of its objective measurements are discussed in Chapter XI of this volume. For educational guidance purposes intelligence tests offer some possibilities, although the evidence is a bit contradictory. School accomplishment in such a specific field as first year algebra can be more reliably prognosticated by a special aptitude test than by an intelligence test.[5] On the other hand, Madsen[6] presents some data (Table 43) which show quite significant differences in the intelligence quotients of pupils passing and failing in certain secondary school subjects.

[5] Greene, H. A., Piper, A. H., *Examiner's Manual* for *Iowa Algebra Aptitude Test.* Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa, 1931.

[6] Madsen, I. N., "The Contributions of Intelligence Tests to Educational Guidance in High School," *School Review,* Vol. 30, 692-701 ; November, 1922.

TABLE 43

MEDIAN INTELLIGENCE QUOTIENTS NECESSARY FOR SUCCESS IN THE SEVERAL SUBJECTS[7]

Subjects	Boys		Girls	
	Failed	Passed	Failed	Passed
Latin I and II	118	135	116	132
Ancient History	105	132	99	127
English	105	129	93	118
Algebra	111	123	99	124
Manual Training		112		
Mechanical Drawing	108	110		
Typewriting	76	109	87	96
Woodworking	70	93		
Telegraphy	76	81		
Household Arts			94	107
Rapid Calculation			66	94
Sewing			76	94
Cooking				83

The data given in Table 44 are probably sufficient to show the relationships which appear to exist between intelligence and vocational success.

TABLE 44

TYPICAL SCORES FOR OCCUPATIONAL GROUPS ON ARMY ALPHA

Approximate Intelligence Levels, Major Occupational Divisions	Median Score	Middle 50 per cent
Unskilled Labor	35	21 – 63
Semi-Skilled Labor	42	23 – 70
Skilled Labor	61	26 – 95
Business and Clerical	96	58 – 145
Professional	140	98 – 184

Occupations are usually arranged in descending order from the professions to unskilled labor on the basis of intelligence.

[7] High school freshmen in Omaha, Nebraska. The intelligence quotients are computed from scores made on the Army Alpha Scale, and are not strictly comparable to intelligence quotients achieved upon the Stanford-Binet.

Within each group, however, mental abilities, while distributed normally, overlap other groups considerably. Consequently, the same individual might succeed in any one of several different types of employment that interested him. Present measures of mental ability provide no adequate index to motivation. Due to the absence of specific measures of interest or "drive" in most intelligence tests and special aptitude tests, the value of the result for guidance purposes is negative rather than positive. The limitations of the individual rather than his possibilities are revealed. Used alone the intelligence test is an adequate guidance instrument, but in combination with other tests or ratings of achievement, interest, specific abilities, and personality traits, it possesses real value for guidance purposes.

Testing for guidance must not be confined to psychological testing alone. While general intelligence tests indicate mental aptitude for scholastic work, educational tests probably still provide the basis for the prediction of achievement on past performance. Educational tests, if sufficiently detailed and analytical, reveal a picture of the strengths and weaknesses in learning as it has been accomplished to date. Such information is valuable for guidance purposes.

Frequently a pupil has no record of previous work in a field related to the one in which he is seeking an occupational outlet. For the purpose of revealing the individual's special aptitudes, tests possessing prognostic power are of great utility. Moss[8] found that scholastic aptitude tests have higher predictive value for medical students than any other single criterion he used. Piper[9] showed that an extensive sampling of roughly scaled arithmetic examples provided the best single index to the pupil's probable success in first year algebra. Similarly Bruce[10] found that scores on a comprehensive test of algebra skills utilized in the solution of problems in plane geometry had the highest prognostic value of any single measure. Experience in the use of Iowa

[8] Moss, F. A., "Aptitude Tests in Selecting Medical Students," *Personnel Journal*, Vol. 10 : 79-94, August, 1931.

[9] Piper, A. H., "The Validity of Certain General and Special Tests for Prognosis in First Year Algebra." Unpublished Master's Thesis, University of Iowa, 1930.

[10] Bruce, H. W., "The Comparative Validity of Certain Measures for Prognosis in Plane Geometry." University of Iowa Unpublished Master's Thesis, Iowa City, Iowa, 1933.

Placement Examinations with college and university students reveals that of all of the tests a combination of the *Mathematics Aptitude Test* and the *English Training Test* gives the best prognosis of probable academic achievement. Thus the need for prognostic instruments in guidance is apparent, even though much of the practical value of these tests does operate negatively rather than positively.

The objective measurement of temperamental traits, those reflectors of the character of the individual, has thus far not reached a very satisfactory state of development, although some progress has been made in recent years. The illustrations presented in the following section give a fair picture of the present status of the material in this field.

III. Measurement of Personality and Character Traits

158. Rating Scales for Personality and Character Traits. The vague qualities of many of the personality and character traits render their objective measurement very difficult. In fact, devices for the measurement of these qualities have generally taken the form of rating scales rather than that of objective tests. However, a number of these devices are so arranged that the individual provides a reasonably objective record of his own reactions. In this respect they resemble tests. Such instruments differ from the typical rating scale also in that the individual responding to the exercises frequently does not know that he is being measured for any particular quality or trait. For example, in the *Bernreuter Personality Inventory*[11] the subject is asked to indicate his reaction to 125 questions by encircling one of the choices which precedes each question. The following samples are typical of the exercises used:

1. Yes No ? Does it make you uncomfortable to be "different" or unconventional ?
2. Yes No ? Do you day-dream frequently ?
5. Yes No ? Do you ever give money to beggars ?

[11] Bernreuter, Robert G., *The Personality Inventory*, Stanford University Press, Stanford University, California, 1931.

15. Yes No ? Do you usually object when a person steps in front of you in a line of people?
25. Yes No ? Do you study the motives of other people carefully?
50. Yes No ? Do you usually try to avoid arguments?
100. Yes No ? Do you prefer to be alone at times of emotional stress?

This arrangement of the material makes it possible to measure several different aspects of personality at one time. According to the author, the scales used in scoring the responses to this test are very reliable. The significance of the individual's response to the questions is brought out by the use of four separate scales in the scoring of answers. For example, Scale B1-N is a measure of neurotic tendency. Persons who score high on this test tend to be emotionally unstable.

In the case of the exercises used in the sample above, a person who answers question 1 with "Yes" scores a plus two points. A "Yes" on question 2 adds five more points. On the other hand a "No" for question 5 deducts six points. When Scale B2-S, the scale for self-sufficiency, is used on these same answers, however, a score of "Yes" for question 1 gives a score of —4; a score of "Yes" on question 2 gives a score of plus 1; a "No" on question 5 gives a score of —3 points; etc. The remaining scales, B3-I for introversion-extroversion, and B4-D for dominance-submission, are applied in a similar manner. Scoring the individual's response to the questions by each of these four scales gives rise to four sets of personality scores for each of which norms are available. Persons scoring high in self-sufficiency are the types who prefer to be alone, do not seek sympathy or encouragement, and tend to follow their own inclinations rather than to seek the advice of others. Persons scoring high on the introversion-extroversion scale are inclined to be imaginative. Those scoring low on this scale rarely worry and prefer to act rather than to dream. Persons scoring high on the dominance-submission scale tend to dominate others in face-to-face situations.

Analysis of personality such as is afforded by the *Bernreuter Personality Inventory* has been used with success and

considerable reliability with high school students, college students and adults. The inventory itself is self-administering. There are no time limits, and each person interprets the questions for himself.

Another type of attempt to secure an unbiased picture of certain personality traits without the subject being completely aware of the traits on which he is to be measured is represented in the *Loofbourow-Keys Personal Index.*[12] This test, according to its manual "is an instrument for the detection of attitudes indicative of problem-behaviour. It is intended for use in group surveys to identify those boys whose personal and social maladjustment is such that they are, or are in danger of becoming serious disciplinary problems." It is standardized for use in the junior high school grades although it has been found useful in senior high school and in continuation school groups. Below are brief samplings from each of the four test-parts comprising the battery.

TEST 1

Directions: This is a test of your word knowledge. Put an X in front of each word you know. There are 100 words.

________ perceive	________ verify
________ restore	________ proportion
________ grole	________ galine
________ luxury	________ exceed
________ rettle	________ patient

TEST 2

Directions: Below are some words and phrases with some statements about each one. Mark an X in front of the one statement under each word or phrase which tells best how you feel about the thing named. Mark only one statement under each one.

(Seventeen items)

1. Chums:

    ______ It is hard to go without them.
    ______ You cannot always trust them.
    ______ They sometimes squeal on you.
    ______ They help you lie out of things.

[12] Loofbourow, Graham C., and Keys, Noel, *Personal Index.* Minneapolis: Educational Test Bureau, Inc., 1933.

3. Teachers:
- ______ They work hard.
- ______ They know they can punish you.
- ______ They are not fair to you.
- ______ They are kind of cranky.

10. Policemen:
- ______ They have it in for the kids.
- ______ They are glad to help you out.
- ______ It is fun to fool them.
- ______ They are just big bluffs.

Test 3

Directions: Read carefully and underline the one response which makes the best answer for you. Underline only one.

(Twenty-one items)

1.	Do you call another person by a nickname he or she does not like?	Almost always	Some-times	Hardly ever
2.	Do you keep right on studying when the teacher goes out of the room?	Almost always	Some-times	Hardly ever
21.	Do you speak pleasantly to all of the people you know even if you do not like them?	Almost always	Some-times	Hardly ever

Test 4

Directions: Answer every question as truthfully and honestly as you can by drawing a line under the right answer, as shown in the samples.

A.	Do you eat more than once a week?	<u>Yes</u>	No
B.	Would you rather have a dime than a dollar?	Yes	<u>No</u>

(Eighty-nine items)

1.	Would you like to wear expensive jewelry, rings, etc.?	Yes	No
2.	Do you feel bored a good deal of the time?	Yes	No
50.	Are you anxious to get away from school and get a job?	Yes	No
89.	Do you know anybody who is trying to do you harm or hurt you?	Yes	No

The test is so constructed that only the undesirable responses are scored. The "problem" responses were determined by comparing the answers given by "problem" boys

with those made by others of the same age and intelligence. Test 1, False Vocabulary, is scored by allowing one point for each fictitious word. The possible score on this part is 30 points. In Test 2, Social Attitudes, three of the four possible answers are socially unacceptable. Each answer so marked counts one error. The score is the number of errors multiplied by two. The possible score is 34 points. Test 3, Virtues, is scored by allowing one point for each fault confessed. The score is obtained by multiplying the number of confessed faults by two. The possible score is 42 points. In the case of Test 4, the Adjustment Questionnaire, the score is the number of "problem" responses. The possible score is 89 points.

The sum of the four scores listed above gives the subject's Personal Index. The highest possible index is 195 points. The reliability of the battery is approximately .90. In the examiner's manual the authors make this statement of the significance of these personal indexes: "An index of 30 or less is clearly insignificant as regards problem behaviour. . . A score of 40 or higher, however, strongly suggests an unwholesome trend, since such scores are made by three out of four reform school boys, as compared with only one in five of the others. Scores of 50 to 60 are much more highly indicative, and occur but rarely in unselected groups. By noting those boys who show high Personal Indexes, say 40 or over, principals, teacher, and counselors will have early brought to their attention those individuals most likely to become disciplinary problems and presumably in gravest need of observation and counsel."

In addition to these two types of more or less objective tests designed to reveal personality and character differences a number of general rating scales are also in common use. These are of two general types. In the first, individuals are ranked according to their standing in regard to the specified character traits, and in the second the character traits are rated and assigned a rank. The best rating scales tend to lessen the spread of teacher estimates and when two or three judgments are averaged are found to give fairly reliable estimates for an individual's traits. On an average the reliability of these ratings is about .55 but varies from .40 to .70

on the best rating scales depending on the traits rated. Whenever possible in using rating scales it is desirable to have two or three teachers rate the same pupil and then average the ratings. Rugg reports this to be a fairly satisfactory method although it is at times difficult to get pupils rated by three different individuals who know the pupil equally well.

Self-rating scales are also used for allowing pupils to rate themselves. These have their chief value in calling the pupil's attention to desirable traits and giving a better understanding of some of the desirable and undesirable traits. It has been found that pupils tend to rate themselves too high on the desirable traits, but in general the reliability of their ratings is about the same as that obtained by other methods.

IV. Measurement of Vocational Interests

One of the very important factors in the vocational success of an individual is his fundamental interest in the field itself. In recent years, considerable attention has been given to this problem of vocational interest by psychologists. As a result a number of helpful devices for measurement of these interests have been developed. Only one will be described and discussed here.

159. **Strong's Vocational Interest Blank.** This *Vocational Interest Blank*[13] devised by Dr. Edward K. Strong of Stanford University has created a large amount of interest and has been used widely in educational and vocational guidance work. The author himself indicates that "it is possible with a fair degree of accuracy to determine by this test whether one would like certain occupations or not. The test is not one of intelligence or school work. It measures the extent to which one's interests agree or disagree with those of successful men in a given profession."

The inquiry blank itself is designed for use throughout the school system. The subject is first confronted with a general inquiry into his social background, his school training, his ambitions for the completion of an educational

[13] Strong, Edward K., *Vocational Interest Blank,* Stanford University Press, Stanford University, California, 1930.

program, his school-subject interests, the field of activity he plans to enter and his reasons for this choice. These general items are supplemented by a more specific inquiry into previous employment.

Part I of the more formal part of the inquiry deals with occupations. The nature of the directions to the subject and the types of items to which his response must be made are indicated in the accompanying brief sampling from this part of the blank. Responses to 100 such items are called for in Part I.

Parts Ia and Ib. ***Occupations.*** Indicate after each occupation listed below whether you would like that kind of work or not. Disregard considerations of salary, social standing, future advancement, etc. Consider only whether you would like to do what is involved in the occupation.

Draw a circle around L if you like that kind of work.

Draw a circle around I if you are indifferent to that kind of work.

Draw a circle around D if you dislike that kind of work.

Work rapidly. Your first impressions are desired here. Answer all the items. Many of the seemingly trivial and irrelevant items are very useful in diagnosing your real attitude.

Actor (not movie)	L	I	D
Advertiser	L	I	D
Architect	L	I	D
Army Officer	L	I	D
Artist	L	I	D
Astronomer	L	I	D
Judge	L	I	D
Labor Arbitrator	L	I	D
Laboratory Technician	L	I	D
Landscape Gardener	L	I	D
Lawyer, Criminal	L	I	D
Lawyer, Corporation	L	I	D
Librarian	L	I	D
Life Insurance Salesman	L	I	D
Locomotive Engineer	L	I	D
Machinist	L	I	D
Undertaker	L	I	D
Watchmaker	L	I	D
Wholesaler	L	I	D
Worker in Y.M.C.A., K. of C., etc...	L	I	D

It is impossible to reproduce here illustrations of the types of reactions called for in the remaining portions of the inquiry

blank. Part II gives an opportunity for the subject to react to a list of 54 amusements. Part III calls for a similar expression of attitude toward a list of 39 school subjects. Part IV deals with attitudes toward various activities while Part V is designed to stimulate reactions to the peculiarities of people. Part VI calls for the individual to select from a list of activities the three which he feels he would enjoy most. Three separate activity lists are provided. In a similar way the subject is asked to indicate in each of these lists the three activities in each group he would like to do least. Part VII provides the basis for a comparison of the subject's relative interest in successive pairs of items. For example, he must indicate whether he would prefer to be a *street-car motorman* or a *street-car conductor*, a *policeman* or a *fireman*, a *chauffeur* or a *chef*; etc. Part VIII calls upon the subject to indicate by means of *Yes No ?* his own rating of his present abilities and characteristics.

From the foregoing description it is apparent that this vocational interest blank provides the basis for a rather searching analysis of an individual's interests. His responses are evaluated and compared with those of successful workers in many different fields of activity with the result that definite interest-leanings may indicate the field in which his success is likely to lie. The authors are convinced that much wasted effort and time on the part of individual students, and much worry and disciplinary difficulty on the part of teachers would be avoided if more attention were given to information of the type afforded by such inquiries into pupils' interests.

V. Measurement of Special Aptitudes

Aptitude is the predisposition of an individual for the development of some special ability or skill. In a certain sense aptitudes are specialized aspects of intelligence, and aptitude tests are measures of special intelligences. Such measures are capable of providing distinctly useful information for the guidance of the individual pupil. In this section space permits the mention of only three types of special aptitudes; namely, mechanical, artistic, and mathematical.

160. Mechanical Aptitude. Mechanical aptitude is the special capacity of the individual to deal successfully with mechanical devices, and to acquire knowledge essential to their selection and operation after suitable training has been given. Individuals with large measures of mechanical aptitude, other things being equal, readily respond to instruction along mechanical lines. On the other hand, individuals with low mechanical abilities are likely to react slowly regardless of the quality of instruction or the opportunities they are given to work with mechanical devices.

It is estimated that at least 40 per cent of the gainfully employed population in the United States is dependent in some measure for its economic success on the possession of mechanical ability. It thus becomes apparent that a knowledge of the student's mechanical ability is important to the teacher from both the guidance and the instructional point of view. Knowledge of the fact that an individual has low or high mechanical aptitude provides an objective basis for guiding him into or out of vocations which involve high degrees of these abilities. Such knowledge is of particular value to trade and continuation school teachers in selecting individuals who are likely to profit by the instruction offered. However, it is well to bear in mind that mechanical ability tests like all other tests must be carefully administered and interpreted, and that at best, they are suggestive but not infallible.

It is widely known that the mechanical ability of different individuals varies widely. It is also known that mechanical ability does not correlate highly with intelligence of the abstract type. The usual relationship is around +.40. This does not mean that individuals with high intelligence as measured by general mental ability tests do not in many cases have high mechanical ability, nor does it mean that individuals with low intelligence always have high mechanical ability. It strongly suggests that there may readily be a concrete aspect of intelligence which is not necessarily an accompaniment of intelligence of the abstract type.

Only one of the several available tests of mechanical ability will be discussed here. This is the *Minnesota Me-*

chanical Ability Tests,[14] which are the outcome of a program of research at the University of Minnesota during the years 1923 to 1927. The tests are quite reliable, and their validity has been carefully checked against objective criteria. The battery of tests includes the following six tests:

(*a*) Minnesota Paper Form Board, Series A and B, for either group of individual testing.
(*b*) Minnesota Spatial Relations Test (Individual Test)
  Boards A and B
  Boards C and D
(*c*) Minnesota Assembly Test (Group or Individual)
  Long form
    Box A, B and C
  Short form
    Box 1 and 2
(*d*) Minnesota Interest Analysis Test
(*e*) Packing Blocks Test
(*f*) Card Sorting Test

TABLE 45
COEFFICIENTS OF RELIABILITY AND VALIDITY ON MINNESOTA MECHANICAL ABILITY TESTS

Test	r_n*	Validity Coefficient (Between test and quality criterion)
Minnesota Assembly, Boxes A, B, C	.94	.55
Minnesota Paper Form Board A and B	.90	.52
Minnesota Spatial Relations Boards A, B, C, D	.84	.53
Packing Blocks	.77	.26
Card Sorting	.90	.19

*$r_{\frac{1}{2}\frac{1}{2}}$ stepped up by Spearman-Brown formula.

The coefficients of reliability and validity shown in Table 45 are those reported by the authors of these tests. The test battery is quite elaborate and rather expensive of time

[14] The materials for these tests may be obtained from the Marietta Apparatus Company, Psychological Equipment, Marietta, Ohio.

and money, but it apparently yields a satisfactorily reliable and valid score for most guidance purposes.

The *Stenquist Assembling Tests*,[15] the *Stenquist Mechanical Aptitude Tests*,[16] the *MacQuarrie Test of Mechanical Ability*,[17] the *Detroit Mechanical Aptitude Examination*,[18] and the *O'Rourke Mechanical Aptitude Test*[19] are among the other better known instruments for forecasting mechanical aptitude.

161. Artistic Talent. *Art Judgment.* The *Meier-Seashore Art Judgment Test*[20] is designed to objectively evaluate certain qualities which appear to underlie artistic accomplishment in the field of art. The test booklets consist of 125 pairs of samples. One specimen in each pair is the work of a master. The other has been changed to vary the composition in some significant factor. The test record sheet on which the subject's responses are recorded specifies the particular factor which is to be considered in reacting to each pair of etchings. The stimuli are arranged approximately in ascending order of difficulty of judgment. The reliability of measurement resulting from the use of the test is high. Suitable norms are available for comparing and interpreting the pupils' scores. The research back of this test indicates that individuals who have the qualities of artistic judgment called for and thus make relatively high scores on the test are quite likely to profit much more significantly from training in art than are those whose scores are low. On the other hand a relatively high score on this test does not necessarily mean that the subject should immediately turn to art as a career for he might have this highly desirable artistic judgment in an adequate amount and yet be handicapped by a lack of manual dexterity or by color-blindness to a point that such a career would be utterly impossible. Such a measure affords an additional cross-section of qualities which are essential to success in this field of activity. Ac-

[15] C. H. Stoelting Company, Chicago, Illinois.
[16] World Book Company, Yonkers, New York.
[17] Southern California School Book Depository, Los Angeles, California.
[18] Public School Publishing Company, Bloomington, Illinois.
[19] Educational and Personnel Publishing Company, Washington, D. C.
[20] Meier, N. C., and Seashore, Carl E., *The Meier-Seashore Art Judgment Test.* Iowa City, Iowa : Bureau of Educational Research and Service, Extension Division, University of Iowa, 1929.

cordingly, it should prove to be a valuable educational guidance instrument.

Musical Talent. The *Seashore Measures of Musical Talent*[21] are widely known musical aptitude tests. They are distinctly prognostic tests and are in no sense measures of musical achievement or appreciation. The tests themselves are on phonograph records and include exercises on basic aspects of pitch, intensity, time, consonance, memory, and rhythm. The tests are valuable in helping to determine those who will probably profit from the study of music. The author of these tests has made a most significant contribution to the science of measurement through his critical analysis of the fundamental factors which make up musical talent.

162. **Aptitude in High School Mathematics.** The *Iowa Algebra Aptitude Test* and the *Iowa Plane Geometry Aptitude Test* are described here merely to illustrate the techniques which may be used in this type of measurement.

The *Iowa Algebra Aptitude Test* is a six-page test folder requiring only 35 minutes for administration. From a large number of possible factors which appear to contribute to successful achievement in first-year algebra, four basic skills have been selected as a result of a very critical statistical analysis of preliminary data. These four abilities deal with (1) arithmetical computations, (2) computations dealing with abstract concepts, (3) the manipulation of numerical series, and (4) the solution of problems involving dependence and variation.

It will be noted that no one of these basic skills involves an acquired algebraic ability as such. Yet the results from the use of this test indicate that students whose scores on the four parts place them at or below the twentieth percentile as given in the norms are almost certain to fail the course or if they do not fail they will practically be carried through the work by the teacher himself. Wherever flexibility in the curricular offerings will permit it such pupils should be diverted into courses in modified mathematics, or out of the

[21] Seashore, C. E., *The Measurement of Musical Talent.* New York: Silver, Burdett and Company, 1929. The test records and the examiner's manual for these tests may be secured through C. H. Stoelting and Company, Chicago, Illinois.

field of mathematics entirely into fields in which they are more likely to succeed. In cases where these adjustments in course requirements can not be made, it is highly desirable for the teacher to discover in advance who the individuals are who are likely to have trouble in mastering the content of the course.

The *Iowa Plane Geometry Aptitude Test*[22] is quite similar in many respects to the *Iowa Algebra Aptitude Test.* It is a four-page test requiring a total of 44 minutes for administration. Part 1 contains exercises in the reading and interpretation of geometric statements and terminology. Part 2 deals with the basic algebraic concepts required in actual work with geometric proofs and problems. Part 3 is made up of simple arithmetical and algebraic reasoning problems. Part 4 contains a modification of the well-known painted-cube test. This portion of the test is designed to measure the individual's power to build up and manipulate a mental image. The statistical evidence obtained in connection with the validation of the test indicates that pupils whose scores place them at or below the twentieth percentile may be counted on with considerable certainty to have continuous trouble with the course if not to fail it completely unless carried by the instructor.

Much remains to be done to provide an adequate basis for effective educational guidance. The tests described in this chapter offer some indication of the service which may be rendered by personality tests, vocational interest charts, character rating scales and prognostic measures of achievement in effecting a better adjustment of the pupil's total personality.

Discussion Exercises

1. Do you accept the idea that in its broader aspects education *is* adjustment? Give your reasons.
2. What are the significant differences between the concepts of adjustment and guidance?
3. Enumerate the guidance values of exploratory courses in the junior high school years.

[22] Greene, H. A., and Bruce, H. W., *Iowa Plane Geometry Aptitude Test.* Iowa City: Bureau of Educational Research and Service. University of Iowa, 1933.

4. In what ways are the adjustment procedures in the general guidance fields distinguished from each other ?
5. From your reading in this field what are the most important character traits ?
6. How are personality and individuality similar and in what specific ways are they quite unlike ?
7. What are the four major personality traits discussed in this chapter ?
8. From the material in this chapter work out an acceptable statement of the functions of educational guidance.
9. Who, why, when or what is the adjusted pupil ?
10. From a guidance or adjustment point of view why are complete cumulative school records a necessity ?
11. Enumerate and evaluate the major objective sources of data for guidance purposes.
12. What is your conclusion with respect to the relative value of general intelligence tests and special aptitude tests for vocational and educational guidance purposes ?

Selected References

Allen, Richard D., *Inor Group-Guidance Series.* Vol. I, "Common Problems in Group Guidance" ; Vol. II, "Case Conference Problems in Group Guidance" ; Vol. III, "Self-Measurement Projects in Group Guidance" ; Vol. IV, "Organization and Supervision of Guidance in Public Education." New York : Inor Publishing Company, Inc., 1934.

Anderson, L. D., "The Relationships of Certain Environmental Factors to Measures of Mechanical Ability," *Yearbook of the National Educational Association, Nature and Nurture,* Part II, 137-150, Bloomington, Illinois : Public School Publishing Company, 1928.

Bawden, W. I., *The Army Trade Test.* Washington, D. C. : U. S. Bureau of Education, Industrial Education Circular, No. 4, April, 1919.

Bernreuter, Robert G., *The Personality Inventory.* Stanford University, California : Stanford University Press, 1931.

Brewer, John M. and Others, *Cases in the Administration of Guidance.* New York : McGraw-Hill Book Company, 1929.

Bruce, H. W., "The Comparative Validity of Certain Measures for Prognosis in Plane Geometry." University of Iowa, Iowa City : Unpublished Master's Thesis, 1933.

Carter, H. D., "The Organization of Mechanical Intelligence," *Ped. Sem. and Journal of Genetic Psychology*, 1928, 35, 270-285.

Edgerton, A. H., *Vocational Guidance and Counseling.* New York : The Macmillan Company, 1926.

Germane, C. E., and Germane, Edith G., *Character Education.* New York : Silver, Burdett and Company, 1929.

Hartshorne, H. A., and May, M. A., *Studies in the Organization of Character.* New York : The Macmillan Company, 1930.

Hull, C. L., *Aptitude Testing.* Yonkers, New York : World Book Company, 1928.

Jones, Arthur J., *Principles of Guidance.* New York : McGraw-Hill Book Company, 1930.

Koos, L. V., and Kefauver, G. N., *Guidance in Secondary Schools.* New York : The Macmillan Company, 1933.

Loofbourow, G. C., and Keys, Noel, *Personal Index.* Minneapolis : Educational Test Bureau, Inc., 1933.

MacQuarrie, T. W., "A Mechanical Ability Test," *Journal of Personnel Research*, Vol. 5 : 329-37, January, 1927.

Meier, Norman C., "A Measure of Art Talent." *Psychological Review Monographs*, December, 1928, pp. 198-199.

Moss, F. A., "Aptitude Tests in Selecting Medical Students," *Personnel Journal*, Vol. 10 : 79-94, August, 1931.

National Association of Secondary School Principals, Bulletin of Guidance in Secondary Schools, Bulletin No. 19, 1928.

Odell, C. W., *Educational Measurement in High School.* Chapter XXII, "Prognosis and Guidance." New York : The Century Company, 1930.

Paterson, D. G., "The Minnesota Mechanical Ability Tests," *Vocational Guidance Bulletin*, Minneapolis Public Schools, Vol. II, No. 9, June, 1928, pp. 1-2.

Piper, A. H., "The Validity of Certain General and Special Tests for Prognosis in First Year Algebra." Unpublished Master's Thesis, University of Iowa, Iowa City, 1930.

Ruch, G. M., and Stoddard, G. D., *Tests and Measurements in High School Instruction.* Yonkers, New York : World Book Company, 1927.

Seashore, C. E., *The Measurement of Musical Talent.* New York : Silver, Burdett and Company, 1929.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York : Silver, Burdett and Company, 1928. Chapter XX, "Prognosis and Special Ability Tests."

Stein, M. L., "A Trial with Criteria of the MacQuarrie Test of Mechanical Ability," *Journal of Applied Psychology*, Vol. 11 : 391-93, October, 1927.

Stenquist, J. L., "Measurements of Mechanical Ability," *Teachers College, Columbia University, Contributions to Education*, No. 130, New York : Bureau of Publications, Teachers College, Columbia University, 1923.

Strong, Edward K., *Vocational Interest Blank*. Stanford University, California : Stanford University Press, 1930.

Symonds, P. M., *Measurement in Secondary Education*. New York : The Macmillan Company, 1927.

Thurstone, L. L., "Mental Tests for Prospective Telegraphers," *The Journal of Applied Psychology*, Vol. 3 : 110-111, June, 1919.

Toops, H. A., *Tests for Vocational Guidance of Children Thirteen to Sixteen*. Teachers College, Columbia University, Contributions to Education, No. 136, New York, 1923.

Watts, F., "Construction of Tests for the Discovery of Vocational Fitness," *Journal of Applied Psychology*, Vol. 5 : 240-52, September, 1921.

Weaver, C. G., "Trade Tests : Their Construction, Use and Possibilities in Industry," *Industrial Arts Magazine*, Vol. 10 : 163-6, May, 1921.

Yerkes, R. N. (Editor), "Intelligence Testing in Occupational Groups," *Memoirs of National Academy of Sciences*, Vol. 15. Washington : Government Printing Office, 1921, Part III, Chapter XV.

Yoakum, C. S., and Yerkes, R. M. (Compilers and Editors), *Army Mental Tests*. New York : Henry Holt and Company, 1920. pp. 196-204.

CHAPTER XIII

MEASUREMENT IN ENGLISH AND RELATED SUBJECTS

This chapter presents a summary of the following points involved in the measurement of English and related subjects:

a. Educational and social significance of English as represented by abilities in composition, language usage, grammar, spelling, and handwriting.
b. Specific aims and outcomes of instruction in English.
c. Specific language skills, oral and written.
d. Measurement of oral English skills.
e. Analysis of mechanics of written composition.
f. Diagnosis and analysis of language usages.
g. Measurement and diagnosis in spelling.
h. Measurement in handwriting.
i. Remedial work in English and related subjects.

I. Aims and Outcomes of Instruction in English

163. **The Educational and Social Importance of English.** Mastery of the tools of expression and communication as represented by such school subjects as composition, language usage, grammar, spelling, handwriting, and reading unquestionably represents the most fundamental of all school outcomes. The fact that instruction in each of these fields begins with the child's first days of school and in some form continues to receive needed emphasis throughout the individual's life may be taken as the best evidences of its educational and social importance. The fact that the individual rarely achieves adequate mastery over these important expressional skills demonstrates their great complexity and difficulty of learning.

English is a distinctly social subject. The language skills function to make communication pleasant and possible. Ac-

cordingly, the activities of the English classroom must provide for social conditions under which communication can take place. The pupil must have something to transmit. He must write or speak to someone for a definite purpose. Whether he wishes to inform, convince, inspire or merely to entertain he must have a reasonable mastery of such tool skills of expression as sentence sense, language usage, mechanics of writing, spelling, reading. He must learn how and where to find the things worth reading and knowing. He must acquire skills which enable him to read with the confident expectation of being himself informed, persuaded, inspired or entertained.

Viewed in this way, English is much more than the specific language skills utilized in recording and transmitting ideas. It involves the skills required for the acquisition of ideas, either through the ear or through the eye, the use of higher mental centers for the critical classification and evaluation of the ideas, and also the more commonly known language skills. In this larger view English ranges the grammatical fields of usage, rules, structure; the mechanical fields of spelling, writing, punctuation, capitalization; the productive fields of speech and reading; the mental fields of organization, imagination, and appreciation. Thus, there are English skills which are definitely receptive and others which are as definitely productive. English delves into many different subject-matter fields since it has no real subject matter of its own. This view of English makes very definite technical demands on the teacher of English, and at the same time assumes an unusually rich cultural background.

164. **Specific Objectives of English Instruction.** Many have attempted to summarize in concise form the important objectives of instruction in English but few have achieved the satisfactory statement presented by Dr. Dora V. Smith.[1] This summary is presented here with the permission of the publishers as a most effective statement of the objectives of instruction in English.

[1] Smith, Dora V., "Diagnosis of Difficulties in English," Chapter XIII of *Educational Diagnosis*, the Thirty-fourth Yearbook of the National Society for the Study of Education. Bloomington, Illinois: Public School Publishing Company, 1935.

Objectives of the Teaching of English, with Their Attendant Contributory Abilities

I. To broaden pupil interests and appreciations through observation and evaluation of experience
   A. To arouse in pupils an interest in, and an understanding of specific experiences which they meet daily
      1. Intimate personal experiences with nature, animals, family, and friends
      2. Larger social experiences of school and community
   B. To assist pupils to recognize the significance of these experiences for the enrichment of personal and social living
II. To develop in pupils the habit of clear, orderly thinking about matters within their own experience
   A. Ability to sense chronological sequence of happenings
   B. Ability to see the relationship of cause and effect in personal experience and in material read
   C. Ability to arrive at trustworthy independent judgments on the basis of fact and experience
III. To promote in pupils the power to organize and express their ideas effectively for others
   A. To assist pupils to gather from experience and from reading, elements bearing upon their immediate purposes in speech or writing
      1. Ability to stick to the subject
      2. Ability to select material of interest to reader or listener
      3. Ability to discriminate between important and unimportant material for the purpose in hand
   B. To help pupils to organize such materials effectively for presentation to others
      1. Ability to marshal evidence or facts concerning the problem under consideration
         a. From observation and experience
         b. From letters or interviews
         c. From books, pamphlets, and magazines
      2. Ability to sense relationships between ideas
      3. Ability to outline ideas in logical sequence, showing power relationship of major and minor aspects through use of recognized outline form
      4. Ability to begin and end effectively
      5. Ability to present a report of reading or experience, keeping the foregoing principles in mind
   C. To give pupils an increasingly adequate vocabulary
      1. Ability to use effectively the general speaking and writing vocabulary of ordinary social and business communication
      2. The desire and power to add to this general vocabulary, words commensurate with individual purposes and ability

3. An interest in words and power of discrimination in choice of words for specific purposes

D. To give pupils adequate mastery of the mechanics of written expression
1. An adequate sentence sense
2. Ability to write well-unified and coherent sentences
3. Adequate skill in the development of ideas in well-rounded paragraphs
4. Mastery of the commonly accepted forms of correct grammatical usage
5. Mastery of the commonly accepted requirements of capitalization, punctuation, and manuscript form
6. Ability in, and sense of responsibility for, the use of legible handwriting
7. Ability in, and a sense of responsibility for, the correct spelling of words in common use

E. To give pupils adequate mastery of the principles of effective speech
1. Sense of purpose and plan in speaking
2. Effective presentation of ideas
3. Audience contact
4. Agreeable, effective use of voice
5. Careful pronunciation, enunciation and articulation

IV. To promote among pupils ability in the clear, simple, effective uses of English necessary to the everyday intercourse of the social and business world
A. Ability to set forth one's own point of view in informal discussion, clearly, effectively, and with proper courtesy
B. Ability to answer directly and exactly a question on which one has the necessary information
C. Ability to communicate necessary facts clearly, forcefully, and with proper conciseness and completeness in telephone conversations, announcements, letters, interviews, reports, instructions or explanations, minutes, etc.
D. Ability to relate an incident or a personal experience in conversation or in writing, making it clear and interesting to listener or reader
E. Ability to carry on conversation with due regard for the interests and rights of others
F. Ability to gather and organize facts or ideas for longer, more formal presentation in speech and writing
G. Ability to conduct or participate in a public meeting or formal discussion with proper attention to parliamentary form

V. To promote the use of correct and effective speech and writing in all subjects and in all activities both within and without school
A. To make pupils conscious of the importance of correct form and usage in the accomplishment of social and business purposes

B. To stimulate in pupils the desire to speak and to write effectively and according to socially acceptable standards of form and usage
C. To develop attitudes of intelligent self-criticism in matters of effective expression and correct form
D. To assist pupils habitually to apply these principles in all speech and writing whether in school or out

VI. To give opportunity for creative expression in the various fields of literary effort
A. To arouse pupils to a sense of the value of originality and creative expression in the everyday intercourse of the social and business world (Activities under IV)
B. To give pupils opportunity to try their powers in voluntary creative activities, such as story-telling, dramatization, informal prose, and the writing of verse

II. Measurement of Oral English

165. Diagnosis of Oral English Disabilities. In view of the very great social importance of the oral language skills it is surprising that more has not been done to perfect standardized measures for this field. Two possibilities of measurement suggest themselves. The first concerns the development of scales for the evaluation of the oral language product. Hitherto the impossibility of securing any kind of objective record of the oral activity of the pupil has made it impracticable to develop such instruments. So far as the writers now know there are no adequate standardized instruments for the measurement of oral composition at the high school level. It appears certain, however, that technical equipment[2] now in use will make it possible to secure recorded specimens of oral compositions which may be scaled in accordance with modern scale construction methods and made available for teaching and testing purposes.

The second of the possible approaches to measurement in the oral English field concerns itself mainly with the analysis of physical disabilities affecting oral language production. Considerable progress in the identification of these specific disabilities and in the development of remedial procedures in oral expression has already been made by inves-

[2] Greene, H. A., Betts, E. A., "A New Technique for the Study of Oral Language Activities." *The Elementary School Journal*, XXXIII, June, 1933.

tigators in the field of speech. It is clear, however, that any classification of speech disorders must necessarily be conditioned by individual points of view. For example, Blanton[3] recognizes four fundamental speech disorders, namely, (1) delayed speech, (2) oral inactivities, (3) letter substitutions, and (4) stuttering. On the other hand, Travis[4] in a much more recent and complete discussion of these problems prefers to group speech disorders under the three following heads:

1. Disorders of rhythm in verbal expression
2. Disorders of articulation and vocalization
3. Disorders of symbolic formulation and expression

Disorders of Rhythm in Verbal Expression. This group of speech disorders includes stammering and stuttering. Travis, after pointing out their basic similarities, states that "stuttering is characterized by the repetition of sounds, words or phrases, while stammering is characterized by speech blocks." Although there are so few serious cases of stuttering that it is not a common pedagogical problem, stuttering has such serious effects on the individual that it is important for teachers to have some notion of the nature and extent of this disorder. Careful surveys of school populations indicate that approximately one pupil per hundred will be a stutterer with the boys far outnumbering the girls in this speech handicap. Since the classroom teacher, no matter how great may be his interest in the nature and the causes of stuttering, can do very little about it, the important thing in connection with instruction in oral language is for him to develop the proper understanding of and sympathy for the stutterer's outlook on life.

Disorders of Articulation and Phonation. Normal speech implies the existence of physical speech equipment adequate to respond to proper stimuli, and of an accurate co-ordination

[3] Blanton, Smiley, *Problems and Methods in the Correction of Defective Speech.* In "Speech Training and Public Speaking for Secondary Schools," a report of a special committee on the National Association of Teachers of Speech, compiled and edited by A. M. Drummond, Chairman of the Committee. The Century Company, New York, 1925.

[4] Travis, Lee Edward, *Speech Pathology.* New York: D. Appleton and Company, 1931. The student will also find a very excellent resume of this author's point of view in *A Handbook of Child Psychology,* Chapter 16. Worcester, Mass.: Clark University Press, 1933.

of the physical and mental aspects of the speech mechanism. Under the category of disorders of articulation and phonation are classified all of the defects which are found in enunciation and voice production, including delayed development of speech.

Travis[5] suggests the following causes of functional speech disorders; namely, faulty training, poor speech sound discrimination ability, poor control of the speech muscles. Furthermore, he points out that in this field there are two types of speech defects, (1) functional defects which are due to bad training, and (2) organic defects which are due to injuries or faulty development of the brain or other organs related to speech. Many of this particular class of speech defects arise from such organic difficulties as abnormal development of the tongue, cleft-palate and harelip, abnormal development of the jaws and teeth, adenoids, defective hearing, etc. The treatment for most of these disorders of articulation and phonation involves medical, mental, hearing, and speech examinations. Since these are generally highly technical in character they should probably be undertaken only by the trained specialist in each field.

Disorders of Symbolic Formulation and Expression. Disorders of symbolic formulation and expression consist essentially of "a lack of power to execute with ease acts connected with articulated speech and the comprehension of spoken words."[6] The location of these defects is largely a clinical rather than a classroom problem. Accordingly, the teacher, upon discovering individual pupils who are unable to articulate or comprehend the spoken word, should immediately refer them to a clinical expert.

From the standpoint of the English teacher and his pupils, speech situations as wholes, rather than separate techniques should be emphasized. The development of technical skills in speech is probably outside of the responsibility of the teacher of English. Aside from the problems of oral language which are peculiarly in the field of speech, a large part of the oral English work should be closely integrated with written English and with literature. The work in oral

[5] *Ibid.*
[6] *Ibid.*

English taught as an integral part of the English course should distinctly avoid the artistic and technical aspects of speech. The teacher's aim should be to conduct the work in such ways that pupils will be equipped to meet effectively the normal speech demands of everyday life. At the beginning of the work, pupils should be led to present and discuss their ideas of the influence of such things as posture, voice, enunciation, pronunciation, facial expression, body adjustments, etc., upon the effectiveness of a speaker. The accompanying outline for an analysis of speech needs has been found very helpful in the classroom.

Outline for Analysis of Speech Needs[7]

1. Does the speaker's position and manner help or hinder his effectiveness?
   - *a.* Does he stand in an easy, natural position (not gripping a table or leaning on it for support, not hiding his hands behind his back or in his pockets, not standing at "military attention")?
   - *b.* Does he give the impression of being alert, interested in what he has to say and in winning the response of his hearers to it? (Does he stand on two feet, not slouching, and speak directly to the audience, not over their heads, not at the floor?) Are his face and body responsive to his thought, alive, and helpful in conveying it?
2. Is the speaker heard distinctly in every part of the room without forcing a loud, unnatural or high-pitched voice?
   - *a.* If he is not heard distinctly, is his enunciation the source of difficulty?
     - (1) Does he fail to pronounce consonants with vigor and precision; are his speech organs lazy and inactive?
     - (2) Are his speech organs tense — his jaw tight, his mouth closed, his throat taut?
   - *b.* If he is not heard distinctly, is the volume of the voice which he uses inadequate, so that his speech is thin and lacking in carrying power?
   - *c.* Does he control his voice volume badly, letting it boom forth at one moment and fade into inaudibility at another?
3. Is the speaker's voice a help to him other than by audibility?
   - *a.* Is the quality pleasant to hear (not nasal, rasping, high-pitched, gruff)?

[7] *Syllabus in English for Secondary Schools* (Grades 7–12), p. 64. Albany, New York: The University of the State of New York Press, 1934.

   b. Does his voice help him by flexibility, modulation as his thought requires, inflection that conveys meaning and prevents monotony?
4. Is his pronunciation in accordance with an acceptable standard?
   a. Has he any faults of pronunciation peculiar to a particular environment?
   b. Has he individual mispronunciations, as inaccurate vowel or consonant sounds, misplaced accent, faulty articulation?

166. **Testing for Speech Defects.** Pupils needing work in speech correction or speech fundamentals should be discovered largely through speech tests given by a special teacher to each entering class. They should be enrolled in special speech classes upon the recommendation of the teacher of English as particular needs are discovered through the regular work in oral English. Only the most carefully trained teacher should attempt the correction of serious speech defects.

III. Measurement of Written English Abilities

167. **Measures of General Merit of Written Composition.** The measurement of general merit of written composition, while dating well back into the history of educational measurement, has not improved in proportion to the attention it has received. Historically, the Hillegas Composition Scale[8] antedates most other attempts to measure educational products. Not only has this scale accomplished much good through stimulating interest in more accurate measurement of written composition, but in its present form as extended and revised by Thorndike[9] it is still a usable instrument.

Among the more useful of the supplements to the Hillegas scale is the *Nassau County Supplement* developed by Trabue for use in the survey of Nassau County, New York. The first seven of the themes comprising this scale were written on the topic, "What I Should Like to Do Next Saturday."

[8] Hillegas, Milo B., "A Scale for the Measurement of Quality in English Composition by Young People." *Teachers College Record,* September, 1912. Bureau of Publications, Teachers College, Columbia University, New York.

[9] Thorndike, E. L., *Thorndike Extension to the Hillegas Scale for the Measurement of Quality in English Composition by Young People.* New York: Bureau of Publications, Teachers College, Columbia University, 1914.

The specimens having scale values of 7.2, 8.0, and 9.0 were selected from compositions published by Thorndike.

The *Willing Scale for Measuring Written Composition* is made up of eight specimens of composition all written on the subject, "An Exciting Experience." Through the definite recognition of the relation of form errors to the general quality of written work, this scale increases its usefulness. Its value is also enhanced through the very clear directions for the collection of compositions for survey purposes. An excellent list of interesting topics is also suggested as the basis for the written work. The use of such standardized lists of topics and the control of conditions under which the writing takes place adds distinctly to the reliability with which written composition abilities may be measured.

The social importance of letter writing as a phase of written language would seem to justify the more extensive use of letter writing scales such as the *Lewis English Composition Scales.* These scales are standardized for use in grades four and above, as are the majority of such scales. They may be used effectively wherever instructional emphasis is given to writing any of the following types: order letters, letters of application, simple narrative social letters, simple expository social letters, or simple narration.

Composition scales are all judgment scales based upon the composite judgments of a group of individuals supposedly qualified to rate compositions on the basis of general merit. The unit of measurement used is a variability unit (either a standard deviation unit or a probable error unit). The differences in quality indicated by the scale values represent positions along a linear scale of general composition merit determined by the proportion of judges who considered a given composition as better than or poorer than another. Thus there is inherent in such scales a certain element of subjectivity due to structure alone. In their use in the measurement of composition quality this is supplemented by the fact that each user of the scale is called upon to make a judgment of the quality of the specimen to be rated in its relation to the specimens comprising the scale. There is evidence that teachers can be trained in the use of the scales

and that the resulting reliability of measurement is definitely increased through training.

In general, any device which will objectify the qualities which enter into general merit of a written composition, tends to make the use of scales easier and more accurate. In the section on the measurement of English composition in the

ANALYSIS OF THE HILLEGAS COMPOSITION SCALE [10]

A. Mechanics (Difficult to Read)	B. Organization (Tiresome to Read)	C. Literary Merit (Interesting to Read)
0–9	30–39	70–79
Meaning uncertain after study	Mere succession of sentences loosely joined	Interesting material marred by imperfect choice of words
10–19	40–49	80–89
Meaning decipherable with difficulty	Disconnected sentences with much irrelevant matter	Well selected material expressed in well chosen words
20–29	50–59	90–99
Meaning not apparent on first reading	Connected sentences; few mistakes; material uninteresting	Exceptional content and quality
	60–69	
	Well organized, but commonplace in content	
SUGGESTIONS: If a composition because of gross errors in mechanics proves difficult to read and understand, it falls in the first division of the scale (0–30). If the meaning is not clear after repeated attempts to decipher the composition, the value assigned should be between 0 and 9 points, depending on the amount which can be read; etc.		

[10] Courtis, S. A., *Gary Survey*. Volume on Measurement of Instruction. New York: General Education Board.

volume of the Gary Survey dealing with the Measurement of Instruction, Courtis suggests an objectification of the Thorndike Extension of the Hillegas Scale which results in distinctly improving the consistency with which relatively untrained and inexperienced judges rate composition samples. Because of the possible contribution it may make in the use of this and other scales, it is reproduced here.

The *Willing Scale for Measuring Written Composition* has accomplished something similar through the use of "Story Value" and "Form Value" scores.

168. **Tests of Grammar and Usage.**[11] While the measurement of the common grammatical usages is not confined to the field of written language the very nature of the subject matter itself makes it necessary to measure it in written form. For those who believe that there is a formal as well as a functional aspect of usage, the *Kirby Grammar Test* should have an appeal. This test is intended to measure the student's ability to recognize the correct reasons for his choice of certain usages. The usage exercises are based upon a composite of studies of the typical errors of children.

Test C of the *Pressey Diagnostic Tests in English Composition* is presented by the aüthors as a grammar test. However, in view of the fact that it confines itself entirely to functional usages it is doubtful if it should be treated as a grammar test as such.

The *Iowa Grammar Information Test* is an attempt to meet the need for a test of the purely informational aspects of English grammar. In addition to its survey use in English classes, it should prove to be a valuable measure of the formal background of grammar needed by students who are beginning the study of a foreign language. By means of eighty objective exercises of the three-answer multiple-choice type it samples into almost all of the commonly-taught phases of English grammar. Two equal and parallel forms are available. Percentile norms are based on 1557 cases in grades seven to twelve.

169. **Form and Usage Tests.** The formal or mechanical aspects of written English are measured in some form by

[11] All tests mentioned in this volume are listed with publishers in the *Appendix*, pages 585 to 595.

most language tests. Only two tests which have this as a specific function are mentioned here. These are the *Briggs English Form Test* and the *Wilson Language Error Test.* The former is a power test designed to measure in a general way the pupil's mastery over certain common elements of written composition. It is standardized for the junior high school grades. Similar in many respects but covering a wider range of material and abilities is the *Wilson Language Error Test.* This useful general survey test, designed for use in grades three to twelve, consists in three stories, each containing approximately twenty-eight common errors made by children in written and oral language usage. It is a proofreading type of test and is limited by considerable inconvenience in scoring as well as by the further fact that frequently children complicate their scores by changing parts of the exercises which are already in correct form.

170. Analytical Measurement of Language Abilities. The abilities which combine to produce skill and efficiency in English are so numerous and so interwoven that it is difficult to conceive of accomplishment in the subject as being represented by a single test score. In fact, almost all of the better and more effective tests in this field are designed to sample into several different aspects of language achievement rather than into one. The very nature of the subject matter makes it doubtful if there is a real diagnostic test in English, but there are undoubtedly several which afford very helpful analytical pictures of achievement in this field. Among these tests suitable for use in the junior or senior high school grades the ones described in the paragraphs immediately following are representative.

The *Columbia Research Bureau English Test* is intended as a measure of general achievement in English covering (1) spelling, (2) the mechanics of English, including punctuation, (3) vocabulary, and (4) literary knowledge. It is designed for use in the upper years of the senior high school and in college classes. In addition to providing a useful basis for maintaining high-school standards in English, it is effectively used for college entrance testing purposes and for placement and guidance purposes. The test is long and thorough, resulting in the high reliability of measurement

which should come from a testing period of 105 minutes.

The *Cross English Test* is designed especially for high-school students and college freshmen. The test is organized in eight parts and covers (1) spelling, (2) pronunciation, (3) sentence structure, (4) punctuation, (5) verb forms, (6) pronoun forms, (7) idiomatic expressions, and (8) miscellaneous faulty expressions. The wide area of skills sampled by this test makes it useful as a final survey of high-school English, and as a college entrance and placement examination.

The *Iowa Elementary Language Tests* are designed for survey purposes in grades four to nine inclusive. The relia-

TABLE 46

IOWA ELEMENTARY LANGUAGE TESTS[12]

(Forms A and B ; Grades 4 to 9)

Test	Number Items	Possible Score	Type of Exercise	Reliability (Av. High-Low)
1. Word Meaning				
A. Synonyms	30	30	MR-4	.95
B. Opposites	30	30	MR-4	.95
2. Language Usage				
A. Correct Word	30	30	Alt.	.90
B. Correct Word Form	30	30	Rec.-Cor.	.86
C. Faulty Expressions	30	30	Alt.	.88
3. Grammatical Form Recognition	30	30	Match.	.91
4. Sentence Sense	30	30	Alt.	.86
5. Sentence Structure	20	40	Br.-Rec.	.71
6. Capitalization-Punctuation				
A. Capitalization	30	30	Rec.-Cor.	.96
B. Punctuation	30	30	Rec.-Cor.	.90
7. Paragraph Organization	14	28	Cross-out	.77
Total	304	338		

[12] Greene, H. A., Ballenger, H. L., *Iowa Elementary Language Tests.* Minneapolis-Philadelphia : Educational Test Bureau, 1929.

bility of measurement on the different parts permits a very useful type of analysis of language limitations. The eleven phases of language ability sampled by this test range over a total of 304 different items with a total possible score of 338 points. In Test 1 which deals with two phases of word meaning the four-answer multiple-choice type of exercise is used. Alternate response exercises are used in two of the three tests measuring phases of language usage. The recognition-correction type is used in Test 2-B, and 6-A and 6-B. A novel type of technique utilizing keyed brackets is used in Test 5.

The *Pressey Diagnostic Tests in English Composition* are useful measures of certain mechanical skills in written language in the junior and senior high school grades. Test A presents twenty-eight exercises in capitalization; Test B calls for responses to thirty punctuation exercises; Test C deals with grammar in thirty exercises of the multiple choice (one out of four) type; and Test D presents twenty-four multiple choice exercises dealing with sentence structure and sentence sense.

Considerable experimentation in determining the best testing technique to use for measuring these mechanical aspects of language indicates that the forms used in these tests are not the most effective and economical. Spaulding[13] has shown that the type of exercises used in Pressey, Test B, Punctuation is not as reliable per unit of testing time and requires considerably longer to score than does an exercise of the recognition-correction type in which a single punctuation skill is measured. This latter type of test exercise is discussed in connection with the *Iowa Elementary Language Tests.* Stickney[14] showed the outstanding superiority of the recognition-correction form of exercise when used to test usages as compared with the four choice multiple choice exercise of the type used in the Pressey tests on grammar and sentence structure. Not only did the recognition-correction type of exercises require much less than half as

[13] Spaulding, E. R., "A Critical Study of Two Methods of Testing Punctuation." Unpublished Master's Thesis, University of Iowa, 1930.

[14] Stickney, George E., "A Comparison of Two Objective Methods of Testing Language Usage." Unpublished Master's Thesis, University of Iowa, 1930.

much time for pupil reaction as the other types, but there was great economy of time in correcting the exercises.

The *Leonard Diagnostic Test in Punctuation and Capitalization* is designed to provide a diagnosis of specific abilities to punctuate and capitalize sentences. It should be of distinct aid to the teacher in identifying pupils in need of individual instruction on these skills, and in measuring individual and class progress in these mechanical phases of English. A detailed study of pupil errors is possible through the use of the diagnostic chart which accompanies the test. The working time varies from 20 to 40 minutes. The reliability coefficients for single grades vary from .87 to .97.

The *Tressler English Minimal Essentials Tests* are designed for grades eight to twelve as well as for college entrance testing. The examination is made up of tests covering (1) the essentials of good usage in grammatical correctness, (2) vocabulary, (3) punctuation, (4) capitalization, (5) sentence structure, (6) sentence sense, (7) inflection and accent, and (8) spelling.

The *Wakefield Diagnostic English Test* is standardized for use in grades nine to twelve inclusive. The testing time is 35 minutes. The test is available in alternate forms. Parts I and II require the naming of the construction of nouns. Parts III and IV concern the recognition of gerunds, participles, and infinitives. Parts V and VI require the identification of transitive and intransitive, passive and active verbs. Parts VII and VIII test a knowledge of the mood of verbs, and Parts IX and X of tense. Parts XI and XII require the student to indicate simple, compound, and complex sentences, and the classification of subordinate clauses into noun, adverbial, and adjective. Parts XIII and XIV involve errors of four types; namely, the dangling participle, the comma blunder, the incomplete sentence, and the rambling sentence. There is a key for rapid scoring. The arrangement of the test makes it very easy to diagnose the weaknesses of individual pupils.

The *Clapp-Young Self-Marking English Test* consists of items dealing with capitalization, punctuation, word form, word usage, and grammar. The test is based on an analysis of text books, courses of study, and studies of errors. There

are two forms to the test, and it is intended for use in grades 8 to 12 inclusive. The time allowed is 25 minutes. The reliability of the test is reported by the authors to be .875. A significant feature of the test is the fact that it is almost self-scoring.

Two English tests are included as a part of the *Iowa Placement Examinations.* Each test is available in two forms, A and B. The tests have proved to be highly reliable, ranging from .87 to .94 for beginning college students. The *English Aptitude Series EA-1, Revised* is made up of four parts as follows:

1. Ability to comprehend and apply rules.
2. Ability to secure correct ideas from English textbook material.
3. Reading comprehension.
4. Composition ability.

The *English Training Series ET-1, Revised* is also made up of four parts:

1. Spelling.
2. Punctuation and sentence structure.
3. Grammar.
4. Clearness, emphasis, force.

The total working time for EA-1 is 43 minutes and for ET-1, 40 minutes. The tests are standardized for college use only.

In addition to the foregoing tests each of which present some analytical possibilities in the measurement of language there are three or four others which should be mentioned. The Pressey tests are designated as diagnostic by their authors but in the light of the criteria for diagnostic measurement which have been set up in this volume, they fall short of being really diagnostic. In fact, it is very doubtful if there are any truly diagnostic tests in the language field. The *Franseen Diagnostic Tests in Language* are diagnostic only to the extent that they identify difficulties dealing with pronouns, verbs, and varied constructions. In spite of their limitations, however, they are very useful tests for survey purposes in grades three to eight. The *New Stanford Achievement Language Test* deals with usage only. The *Unit Scales of Attainment* in language are considerably

more comprehensive, dealing with three aspects of language ability, capitalization, punctuation, and usage.

171. **English Diction Test.** The *Shutte English Diction Test* is composed of a fourteen page narrative in which a total of 410 words, phrases, and other expressions are arranged for easy identification. The pupil is instructed to check each word or group in which there is an error in English diction.

After the test has been given, the test booklets are returned to the pupils, and each pupil is then provided with a manual which discusses the various types of errors included in the test.

The test has value in that (a) it points out specific errors in the pupil's usage, which demand attention; (b) it shows the student certain faults which he must correct; (c) it provides examples of both good and bad English in a narrative which the pupil can study with the aid of the manual and in class discussion, in order to improve his own English. It points to the building of right habits of usage rather than to the learning of rules.

IV. Spelling as a Factor in English Achievement

Spelling is a tool which is essential to all writing, and as such must be constantly emphasized and correlated with all phases of instruction in English. As a matter of fact, since it is impossible for the teacher of English to teach the correct spelling of many words found in such subjects as science and mathematics, all high-school teachers must assume responsibility for the correct spelling of the technical words in their own fields. Spelling is one of the subjects in which learning is specific. A pupil does not just learn spelling, but he learns to spell specific words. He may acquire a general method of learning to spell, but the words he learns to spell are mastered as a result of a definite application of effort and attention.

172. **Measurable Qualities in Spelling.** Systematic instruction in spelling is usually not a part of the high school program of studies. Presumably the pupil has acquired a

method of learning to spell as a part of his elementary school training. Presumably also he has been brought into contact with fairly extensive vocabularies of socially useful words in these earlier school experiences. There is considerable evidence, however, that in many high schools specific emphasis on spelling may not be entirely wasted effort.

Ability to spell is usually revealed by the pupil's success in writing in either list or context form words which have been selected from some vocabulary list of social or school importance. For more mature students of the high-school age the ability to recognize the correct or incorrect spellings of words is a most desirable skill. This type of ability is most commonly measured by tests of the proof-reading type in which the pupil identifies the misspelled word and corrects it.

173. **Construction of Spelling Tests.** In constructing spelling tests there are at least four problems which require careful consideration. These are discussed briefly in turn.

What Words. In the construction of a spelling test one of the first problems is that of selecting the words to be included. The values of spelling are almost entirely specific and lie in the ability of the pupil to spell words that are quite certain to be used. It is important, therefore, that the content for a test should be sampled from those words that are and will be ultimately of maximal usefulness to the pupil. Theoretically, such word-lists may be obtained from analysis of the writing vocabularies of children, the writing vocabularies of adults, or from words common to both.

There are numerous valuable sources of word lists for spelling teaching and testing purposes. Historically the *Ayres Spelling Scale*[15] was the first serious attempt to develop a socially evaluated list of words and to place them in scaled order of spelling difficulty. The list itself was secured by sampling the first word in each line of two thousand short personal and business letters. The total number of running words analyzed was only slightly over twenty-three thousand, while a total of 2001 different words was

[15] Ayres, Leonard P., *Ayres Spelling Scale.* The Russell Sage Foundation, New York, 1915.

found. One thousand of these words were submitted to children in different grades in the form of list tests. On the basis of approximately two hundred spellings per word these words were placed in a scale in accordance with their spelling difficulty for the elementary school grades. Later Buckingham[16] extended the Ayres Scale by the addition of 505 new words selected for their difficulty mainly from lists found in a number of spelling books. In 1917 a study by Dr. W. N. Anderson[17] opened the way for a more extensive series of scales on a vocabulary of social importance. On the basis of 2977 words found by Anderson, Ashbaugh in 1919 prepared the *Iowa Spelling Scales.*[18] Separate scales are provided for each of grades two to eight. The words are arranged in twenty-five groups, each group varying from the one above and below by equal differences in difficulty.[19]

Two comprehensive vocabulary studies containing extensive lists of unscaled words also afford valuable sources of spelling lists. These are the Thorndike *Teacher's Word Book*,[20] and the Horn *Basic Writing Vocabulary.*[21] The *Teacher's Word Book* is a list of ten thousand words which were found to occur most frequently in a count of several million words taken from many sources. It is both a reading and a writing vocabulary. The *Basic Writing Vocabulary* is a list of ten thousand words chosen from varied types of adult writing. The words are classified on the basis of frequency, and each word frequency is compared with that given in other vocabulary studies. This study took into account all previous spelling vocabularies, and as a result has greatly influenced the content of recent spelling text-

[16] Buckingham, B. R., *The Buckingham Extension of the Ayres Scale.* Bloomington, Illinois : The Public School Publishing Company, 1918.

[17] Anderson, W. N., *Determination of a Spelling Vocabulary Based upon Written Correspondence.* State University of Iowa, Studies in Education, Vol. II ; No. 1. Iowa City, 1917.

[18] Ashbaugh, E. J., *The Iowa Spelling Scales.* Bloomington, Illinois : Public School Publishing Company, 1919.

[19] Spelling difficulty must be understood here to represent the per cent of misspelling of a word in a given grade without regard to, or any information about, the previous opportunity for learning the word the pupils may have had. The quality represented is really "persistence of error" and not innate "learning difficulty."

[20] Thorndike, E. L., *The Teacher's Word Book.* New York : Teachers College, Columbia University, 1921.

[21] Horn, Ernest, *Basic Writing Vocabulary.* Iowa City : University of Iowa. April, 1926.

books and tests. Only 3009 of the ten thousand words in this writing vocabulary have been designated by the author as basic for elementary school spelling lists. Naturally the more difficult words for high school students are found in the lower social frequencies.

Of these several sources of data on the spelling difficulties of words the *Iowa Spelling Scales* represent very useful elementary school lists. The words are of known social significance, and of known difficulty. Furthermore the lists are long enough to afford a reasonable basis for sampling. The *Ayres Spelling Scales* are useful but are limited to only a one thousand word sampling. The *Buckingham Extension of the Ayres Scale* does not add much to the available sources of valid spelling material since the words added to the original Ayres scale are of unknown social origin and are included merely to add to the difficulty of the list.

The most satisfactory sources of evaluated word lists for high-school spelling testing and instructional purposes are the so-called *Seven S Spelling Scales* (*Sixteen Spelling Scales Standardized in Sentences for Secondary Schools*, hence, the *Seven S Scales*), and the *Simmons-Bixler Standard High School Spelling Scales.* The former consists in sixteen scaled lists of sentences in which the significant words are embedded. The first twelve in these test lists are equivalent in difficulty as determined by the per cent of misspellings. The remaining four scales of the series are more difficult. The words comprising the scales were taken from the second and third thousand most commonly used words as determined by a composite of numerous studies. The large number of forms makes it possible to test frequently during the high school course without repeating the same form of the scales. The tests are intended for use in grades seven to twelve. The time required is approximately ten minutes per scale. According to the authors the tests are sufficiently reliable to determine the spelling ability of a group by the use of one scale, and the ability of an individual by the use of three scales.

The *Simmons-Bixler Standard High School Spelling Scale* contains words of relatively high social frequency selected from the *Commonwealth List* as suitable for use in

the junior and senior high school grades. The scale is in two parts. Part I contains sixty-four lessons of forty words each (2560 words) for study purposes. Percentage norms for grades seven to twelve inclusive are given for these lessons. In part II 2910 words are arranged in alphabetical order. For each word the percentage of spelling accuracy for grades nine to twelve is shown. The first part thus provides a helpful study list while the second part makes it easy to select word-lists of specified spelling difficulty for instructional or testing purposes.

How Difficult Words. It is well known that some words are more difficult to spell than others, i.e., some words are more frequently misspelled than others. If words are selected at random from any one of the lists indicated above, some of them will be easy and some relatively difficult. If equal credit were given for each word thus selected, an error would be introduced into the estimate of the pupil's spelling efficiency. Words for a test should be selected in terms of their known difficulty. The words comprising the *Ayres Spelling Scale*, the *Iowa Spelling Scales*, and the *Simmons-Bixler Scales* have been so classified by having the words spelled by large numbers of children, and the relative difficulty of each word determined by the per cent of correct spellings of each word. The words to be included in the test for any grade should be adapted if possible to the ability of the group to be tested. Classes of average ability appear to respond best to words of approximately 50% difficulty.[22] On the other hand, if the test is to be given over a wide spread of ability, words ranging from fourteen to eighty-six per cent standard accuracy with a mean of fifty per cent tend to give a distribution more closely approximating the normal frequency curve with the pupils grouped more closely around the mean. In general, it is probably safe to say that the words to be included in a test for any grade should be those on which there are from 40 to 70 per cent misspellings. Tests made up of such words will give a reliable measure of spelling ability, since the words will not be so

[22] Cook, Walter W., *The Measurement of General Spelling Ability Involving Controlled Comparisons Between Techniques.* University of Iowa Studies in Education, Volume VI, Number 6. Iowa City : University of Iowa, February 15, 1932.

easy that there will be many perfect scores or so difficult that there will be many low scores.

How Many Words. The purpose the test is to serve will determine the number of words to use. While 25 words is possibly a sufficient number for survey purposes a larger number is needed to reveal the spelling ability of individual pupils. For individual testing purposes Otis recommends the use of lists composed of at least 100 words, while Starch recommends the use of 200 words. Morrison and McCall[23] have included only 50 words in each of their tests. On the whole it appears that a minimum of one hundred words should probably be used for individual testing purposes in spelling. Possibly fifty words is not too many to use for the measurement of general class accomplishment.

How Given. The question of the form in which spelling words should be presented for testing purposes has called forth much debate in the past. The problem has also been subjected to experimental study with results which are not too conclusive when considered in the light of practical classroom procedures. Spelling words do not appear in list or column form in any life situations. The actual need for spelling arises only in writing connected discourses in which one must choose his words and write them without having to devote much thought to correctness in spelling. Many educators, therefore, have advocated teaching and testing spelling in contextual form.

The dictation of spelling exercises in context does conform somewhat to the conditions of written composition. Nevertheless, the pupil is relieved of the necessity of fixing his attention upon the organization of ideas and the choice of appropriate vocabulary, both of which are most significant in writing. While there is not entire agreement on the point, evidence[24, 25] tends to show that the column method

[23] *Morrison and McCall Spelling Scale.* Composed of eight lists of 50 words each, all lists being of equal difficulty. The lists are based upon the Ayres Spelling Scale, the Buckingham Extension, and the Thorndike Teacher's Word Book. Useful for testing in grades 2-9. Published by Public School Publishing Company, Bloomington, Illinois.

[24] McKee, P., "Teaching and Testing Spelling by Column and Context Forms." *Fourth Yearbook, Department of Superintendence,* 1926.

[25] Hawley, W. E., and Gallup, Jackson, "The 'List' Versus the 'Sentence' Method of Teaching Spelling." *Journal of Educational Research,* Vol. 5, April, 1922.

is fully as effective as the context method in most teaching and testing situations in spelling. In many respects, the time factor definitely favors the column method, since many more words can be spelled in column form than in context in a given period of time.

No system of teaching, however, should ignore the spelling of words in context. A pupil's achievement in spelling should take into account his spelling in all normal writing activities in other school subjects. From the standpoint of its practical application in the classroom this means that probably a combination of the list and the context dictation exercise will provide the most effective instrument for measuring spelling ability for survey purposes.

The list test for dictation to a class may be made up of approximately fifty words ranging in difficulty from perhaps forty up to seventy per cent. The rate at which these words should be dictated can not be given for all grades, but Cook[26] has given some data which may be suggestive. Eighth grade pupils are best satisfied with a list test when fifty per cent standard accuracy words are dictated at the rate of one word every eleven seconds. In higher grades this rate of dictation may be slightly increased.

In context dictation exercises the critical words should not be placed at the end of the sentence since the pupil may have difficulty in understanding the first word or may not be able to complete the writing of the last word within the time limit. The exercises should be dictated at a rate approximating the standard rate of writing for the grade in which the test is to be used. The rate of dictation is dependent upon (1) the rate at which students of the grade normally write, (2) the number of letters in the exercises to be written, and (3) the time required for the examiner to dictate the exercises. Experience has shown that approximately ten per cent of the writing time for the sentence should be added to compensate for the time required for dictation. The accompanying table, which is based upon Freeman's rate of writing standards, gives the total time allowance in seconds per letter for the different school grades.

[26] *Ibid.*

These rate standards are adjusted to provide for dictation time.

TABLE 47

RATE STANDARDS FOR USE IN PREPARING TIMED DICTATION EXERCISES

	Grade						
	2	3	4	5	6	7	8
Time allowance in seconds per letter..........	1.84	1.38	1.18	1.01	.92	.83	.73

The use of this table is demonstrated in connection with the preparation of the sample dictation and list test shown on page 336. This 50-word test requires approximately ten minutes for administration. The first sentence in this test, "He is *evidently entitled* to it," is made up of 25 letters. Since this test is designed for use in Grade 7 the time allowance for seventh grade as given in Table 47 is used. The number of letters (25) should then be multiplied by the rate in seconds (.83) which gives 20.75 seconds, the time required for a seventh grade pupil to write twenty-five letters plus the time required for the dictation of the sentence. In round numbers this would be considered as 21 seconds.

In the administration of timed dictation tests the usual practice is to start the test with the second hand of the watch at 60. Read the first sentence. Then wait until the second hand reaches the time indicated at the left of the next sentence and then read the second sentence, etc. Read each sentence once only. In beginning the test do not suggest to the children that it is a spelling test. Something like the following is a good approach: "I am going to read to you some sentences which I wish you to write down; the sentences are not long and you need not hurry. Listen carefully and then write each sentence just as I have read it. Be careful of your writing and spelling."

174. **Diagnosis of Spelling Disabilities.** Spelling tests and scales afford useful sources of material which may be used

to determine both the student's present status in spelling and his growth in accomplishment as a result of a period of instruction. Scales provide the most effective materials for identifying the spelling difficulties of individual pupils. Samplings from scales used as tests give the teacher an objective basis for the study of these personal difficulties through the accumulation of individual lists of words which are sources of trouble to the student.

Specimen 50-Word Spelling Test for Grade 7

Dictation Exercise

Start with watch at	
60	He is *evidently entitled* to it.
21	Write a *brief description* of a house.
46	We *prefer* this *proposition* ourselves.
13	We have an *excellent commercial* club here.
42	We are *studying science* and *literature* today.
14	The city *council* will not *interfere* with the work.
48	We *acquire literary appreciation* through study.
20	The *annual agricultural demonstration* was held.
55	The *university* is *altogether* too *crowded* this year.
31	The *superintendent authorized athletics* in the school.
Stop — 9	

Now have the children write the following words on the blank side of the paper. Pronounce these as you would an ordinary spelling lesson at the rate of about one word each 12 seconds.

Spelling List

alfalfa	expensive	instrument	catalogue	probably
additional	February	absolutely	disappoint	enormous
confidence	genuine	ninth	correspond	relieved
established	gradually	advisable	hastily	legislation
disposal	imagine	capacity	individual	naturally

Cumulative Personal Lists. To a large extent remedial procedures in spelling should be undertaken directly in connection with teaching. The words misspelled by pupils in their daily work and tests are obviously the words to which they should give special attention. Each pupil should be encouraged to keep an individual list of such words and should be stimulated to master them. Occasional spelling

periods should be put aside for studying and testing these individual lists. If such lists are properly utilized each student will come to regard his individual "dēmon" list as a means for eliminating spelling deficiencies rather than as a form of punishment for misspellings.

Analysis of Written Work. Pupils' written work in all subjects should be carefully checked for spelling errors which every pupil should keep in a list. The pupil should realize that he is responsible for the mastery of these troublesome words. The teacher must see to it that pupils develop a sense of pride in correct spelling and dissatisfaction over incorrect spelling. If this is accomplished, the majority of pupils will soon assume a new responsibility for spelling correctly the words used in their written work. They will check their spelling before turning in their papers. They will habitually appeal to authority when in doubt, which implies the development of desirable habits of using the dictionary.

Individual Pupil Diagnosis. The following items of information procurable through observation and measurement are invaluable in diagnosing individual pupil disabilities and should be used as much as possible in connection with the analysis of the pupils' spelling habits:

1. Intelligence quotient
2. Spelling grades
3. Reading grades
4. Writing grades
5. Attendance data
6. Visual defects, data
7. Auditory defects, data
8. Speech data
9. General health data
10. Personality characteristics; industry, aggressiveness, independence, attentiveness, exactness

Many investigators of spelling disabilities[27, 28, 29], have abandoned the procedure of deducing the causes of spelling difficulties from an analysis of errors and are now devoting

[27] Hollingworth, Leta S., *Psychology of Special Disabilities in Spelling.* Teachers College Contributions to Education, Columbia University, No. 88.

[28] Gates, A. I., *The Psychology of Reading and Spelling.* Teachers College Contributions to Education, Columbia University.

[29] Witty, P. A., "Diagnosis and Treatment of Poor Spellers." *Journal of Educational Research,* Vol. 13, January, 1926.

their time and energies to studying the work habits of pupils by means of careful observation and tests. One of the chief causes of poor spelling achievement lies in the failure of the student through no fault of his own to utilize an adequate method of study. Poor study technique upon the part of the pupil may be detected through careful observation and testing.

Another significant cause of spelling disability is imperfect pronunciation and articulation of words. The contributing factors may be imperfect hearing or improper pronunciation and enunciation of the words by parents, teachers, and pupils. Frequently words are misspelled because they are spelled as they are heard or pronounced. It is interesting in this connection to note that improper pronunciation by the pupil himself is even more likely to cause errors in spelling than such mispronunciation by the teacher in presenting the word either in teaching or testing. Children should pronounce the words to themselves in learning to spell them. It is therefore very important that they form proper habits of pronunciation and enunciation.

V. Handwriting as a Factor in Language Ability

175. **Measurable Qualities in Handwriting.** The measurement of handwriting quality is concerned with two factors, (1) quality, or degree of legibility, and (2) speed, or the quantity of writing produced in a given unit of time.

Quality. In modern practice the quality of handwriting is usually determined by comparing a sample of the pupil's handwriting with specimens in a standard scale. While this method of evaluating handwriting specimens is somewhat subjective, experience shows that considerable skill and objectivity can be developed through training in the use of such scales. At one time the measurement of handwriting simply involved the comparison of the script produced with the copybook sample. This resulted in over-emphasis on the shape and shading of letters and in the formation of beautifully engraved lines. Rate and quality were not the objectives of writing instruction or of measurement under those conditions.

When a pupil's handwriting quality is found to be unsatisfactory good teaching demands that so far as possible the reasons for it must be determined. Freeman has identified the elements of handwriting quality with unusual success and has incorporated them in his *Diagnostic Handwriting Chart.*[30] The essentials of quality in writing are measurable within reasonable limits. A number of scales have been developed for use in measuring quality, but they differ greatly in the number of elements of quality measured and in the numerical designation of each quality difference. Therefore, it is difficult to compare the results secured from the use of one scale with those secured from another. The accompanying table summarizes norms of handwriting quality and rate derived from a number of different scales.

TABLE 48

GRADE STANDARDS IN WRITING [31]

Grade	Ayres Scale	Form or Quality Thorndike Scale	Freeman Scale	Speed Letters per Minute
II	35	8.5	11	30
III	39	9.3	12.5	44
IV	46	10.2	14.5	51
V	50	11.0	16	60
VI	57	11.9	18	63
VII	62	12.7	20	68
VIII	66	13.5	21	73

The standards for quality as shown in Table 48 are considered by many to be sufficiently high. The quality values assigned for the various grades are based upon the median performance of many school children. Therefore, these values indicate the quality of handwriting which exists; they do not necessarily indicate the quality which should be ob-

[30] Freeman, F. N., *Chart for Diagnosing Faults in Handwriting.* Boston: Houghton Mifflin Company, 1914.

[31] Freeman, F. N., *Eighteenth Yearbook of the National Society for the Study of Education,* Part II. Bloomington, Illinois: Public School Publishing Company, 1919.

tained. In an effort to set a social standard of handwriting quality Koos[32] checked the quality of handwriting demanded in social correspondence and in certain vocations. He found that the average quality of 1,053 specimens of social correspondence of adults was 49.5 on the Ayres Scale, which is equivalent to the median for the fourth grade. He then asked 286 persons to indicate the quality of writing they considered adequate for general purposes. The majority of them (4/5) considered the quality 60 (Ayres Scale) as adequate; over 50% of them considered quality 50 to be adequate. The quality of writing adequate for vocations was determined by scoring the writing specimens of 1,127 persons in 24 different vocations. In summarizing his analysis of these samples of handwriting, Koos concluded: "The fact that some (pupils) will go into pursuits demanding a quality better than 60 should not be offered as a justification for requiring all pupils to attain that better quality. . . Since all should be required to learn to write as well as 60 for purely social use, to train pupils to write this quality is the task of general education; to teach some who are going into commercial or other vocations requiring a higher quality . . . to write this better quality is the task not of general but of vocational education." He also makes the following statement: "In the light of these facts it is difficult to see why . . . a pupil should be required to spend the time necessary to learn to write better than the quality of 60. There is even considerable justification for setting the ultimate standard at 50."

Rate. The rate at which pupils write is of considerable importance. The person who is able to write more rapidly than others and with approximately the same quality has an obvious advantage in the field of written expression, provided, of course, that ideas come to him as rapidly as he is able to transcribe them. The measurement of rate in writing is much less difficult than measurement of quality. Rate of writing can be measured most conveniently by asking pupils to write, within carefully controlled time limits, selections from standardized copy. If the pupils all write from the same selection and if they have all thoroughly memorized

[32] Koos, L. V., "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools." *Elementary School Journal*, Vol. 18, pp. 423-446.

it, the number of letters each pupil writes in the time allotted can easily be computed as the pupil's rate score. Table 48 indicates that students at the end of the eighth grade write at the rate of 73 letters per minute.

176. **Methods of Measurement.** Measurement of the quality of handwriting and the rate at which it is produced is accomplished by the evaluation of handwriting specimens secured under standard conditions. Accordingly the first step in the process of measuring handwriting is that of securing these specimens under controlled conditions.

Securing Handwriting Specimens. Three factors appear to affect the conditions under which handwriting specimens for scaling are secured. The character of the copy which the pupils are called upon to write may significantly influence their reactions. Whatever sample is used should be simple and easily understood so that the pupils will not be unduly affected by spelling and vocabulary difficulties. The instructions which are given to the pupils may also influence the quality and rate of their writing. Therefore, care should be exercised to use very precise directions such as those suggested with the handwriting scale in use. If the specimens are to be used as a survey test, standard copy such as the "Gettysburg Speech" should be used. The time allowance for writing the specimens is a third factor which must be considered in the collection of writing specimens. In the standardization of this scale, Ayres used the first four sentences of Lincoln's "Gettysburg Speech" and allowed each child two minutes in which to copy as much of this material as possible. Since then it has become a rather typical practice to allow a two-minute period for the writing of such samples.

Securing Rate Scores. Rate of handwriting is expressed in terms of the number of letters written per minute. This is determined by counting the total number of letters written by each pupil and dividing this by the number of minutes the pupils were allowed to write.

Securing Quality Scores. The quality of the handwriting specimen which is being measured is determined by moving it along the scale until a specimen is found which closely matches it in quality. The quality value of the scale sample

is then assigned as the quality of the sample of the pupil's handwriting.

177. **Accuracy in Measurement of Handwriting.** Skill in the evaluation of handwriting specimens requires a thorough understanding of the scale to be used. It is desirable therefore for the teacher, prior to any attempt to use the scale for the measurement of handwriting quality, to study carefully the scale itself, the directions for its use, the norms and the specific functions which the particular scale is expected to perform. The accurate and reasonably objective rating of handwriting samples on a scale requires considerable skill which experience shows can be developed through practice. For this purpose standard sets of writing samples of known quality such as those prepared by Thorndike [33] and by Shaw [34] are very useful. Hurt [35] found that by using practice sets of this character, "seventeen of twenty-one judges were able to reduce their errors to within negligible limits."

178. **Merit Scales.** Handwriting scales may be divided into two groups — (1) general merit scales and (2) analytical and diagnostic charts and scales. The choice of a scale depends upon the purpose it is to serve.

The Thorndike Scale [36] was the first writing scale to be devised. This scale is designed for grades V to VIII inclusive and consists of a series of handwriting specimens so arranged that they increase in order of merit from a quality of 4 units above zero to one of 18. Its purpose is to aid teachers in grading handwriting for "general merit" on the basis of three characteristics : beauty, legibility, and character.

The *Ayres Handwriting Scale* [37] was the next scale to be devised. It differs from the *Thorndike Scale* in that the unit of measurement used was legibility, determined by the speed and ease with which samples of handwriting were read by a number of trained and competent judges.

The specimens of handwriting for both the Thorndike and

[33] Thorndike, E. L., "Teachers' Estimates of Specimens of Handwriting." *Teachers College Record,* Vol. 15, No. 5, Columbia University.

[34] Shaw, Lena, *Supervision of Handwriting.* Detroit, Michigan.

[35] Hurt, A. O., *A Study of Reliability of the Thorndike Handwriting Scale.* Unpublished Master's Thesis, University of Missouri.

[36] Thorndike, E. L., "Handwriting." *Teachers College Record,* Vol. 2, 1910.

[37] Ayres, Leonard P., *Scale for Measuring the Quality of Handwriting of School Children.* Bulletin No. 113, Division of Education, Russell Sage Foundation.

Ayres Scales, while presumably sampled from all levels of the school, were taken largely from the upper grade levels. This presumably should increase the suitability of these scales for use in the upper grades and high school.

The *American Handwriting Scale*[38] developed by Paul V. West is one of the most recent and comprehensive of the general merit scales. Among a number of distinctive features are at least two which deserve special mention: (1) A separate scale is provided for each grade from two to eight; (2) The samples have been scaled for both quality and rate, the poorer samples being those written at the slower rate and the better samples being the ones written at the more rapid rate. The existence of the separate scales for grades two to eight inclusive permits a somewhat more exact evaluation of quality of writing in its relation to grade location.

179. **Diagnosis of Handwriting Faults.** Improvement in handwriting instruction depends to a large degree upon the teacher's knowledge of the elements which make for quality in the product, and the use of instruments which are adequate to reveal significant differences in quality. Inferior products of handwriting instruction may be due to lack of skill or mastery in many different phases of the writing act. Freeman's *Chart for Diagnosing Faults in Handwriting*[39] meets this need for securing separate measures of the several aspects of handwriting performance. This scale may be used to measure the whole class, but is most effective when used to diagnose the writing of pupils who rank conspicuously below the grade norm as revealed through use of some general merit scale. Freeman's Chart identifies deficiencies in the following five characteristics of handwriting: (1) uniformity of slant, (2) uniformity of alignment, (3) equality of line, (4) letter formation, and (5) spacing.

The following list of handwriting defects and their causes should be useful to the classroom teacher. These items are reproduced from Freeman's *The Teaching of Handwriting*.[40]

180. **Remedial Work in Handwriting.** Remedial work in handwriting, as in every other tool subject, must be adapted

38 Published and copyrighted by A. N. Palmer Company, 1929.

39 *Ibid.*

40 Houghton Mifflin Company, Boston.

to the specific needs of the individual pupil. Differences in the writing ability of pupils are enormous. Remedial instruction can be effectively applied only after careful diagnosis of the individual pupil. The problems of corrective teaching in handwriting are quite technical, and if adequately treated should occupy a volume in themselves. In this discussion only a few brief suggestions can be made. In a

ANALYSIS OF DEFECTS IN WRITING AND THEIR CAUSES

Defect		*Causes*
1. Too much slant.........	(1)	Writing arm too near body
	(2)	Thumb too stiff
	(3)	Point of nib too far from fingers
	(4)	Paper in wrong position
	(5)	Stroke in wrong direction
2. Writing too straight......	(1)	Arm too far from body
	(2)	Fingers too near nib
	(3)	Index finger alone guiding pen
	(4)	Incorrect position of paper
3. Writing too heavy.......	(1)	Index finger pressing too heavily
	(2)	Using wrong pen
	(3)	Penholder too small diameter
4. Writing too light........	(1)	Pen held too obliquely or too straight
	(2)	Eyelet of pen turned side
	(3)	Penholder too large diameter
5. Writing too angular......	(1)	Thumb too stiff
	(2)	Penholder too lightly held
	(3)	Movement too slow
6. Writing too irregular.....	(1)	Lack of freedom of movement
	(2)	Movement of hand too slow
	(3)	Pen gripping
	(4)	Incorrect or uncomfortable position
7. Spacing too wide........	(1)	Pen progresses too fast to right
	(2)	Too much lateral movement.

general way, improvement in handwriting appears to be dependent upon (1) adequate physical conditions and materials, and (2) the maintenance of the proper psychological conditions for effective learning.

Physical Conditions and Materials. Prominent among the physical factors affecting the pupil's handwriting is the matter of his desk. The pupil's desk should be adjusted to his height so that when he is seated normally his thigh is

at right angles to the lower part of his leg and his feet flat on the floor. In accordance with most modern methods of writing the pupil's body when writing should face squarely the middle of the desk and bend slightly forward at the hips. Both forearms should be well up on the desk, the left holding the paper, the right wrist raised and inclined slightly to the right. It is necessary that the pupil be taught to move the paper upward and to the left as the writing progresses. The shifting is done with the left hand, while the right arm is held in the correct position. There is some difference of opinon as to the best position of the writing arm. It is generally agreed, however, that the writing hand should be supported on the third and fourth fingers and that the wrist should not be tilted more than 45 degrees. The forearm of the right hand should be perpendicular to the line of writing. The pen should be grasped lightly and in such a way that the forefinger is below the thumb and at least one inch above the point of the pen.

Experiment and observation show that writing is a combination of whole-arm, forearm, wrist, and finger movements. It is not possible to eliminate entirely finger movement, even in so-called "muscular movement writing." Freeman says, "good writers use more arm movement than poor writers, although finger movement is also employed by the best writers." [41]

Psychological Conditions. Next in importance in preparing the way for effective mastery of writing faults is the provision of desirable psychological conditions. The establishment of a desire for improvement on the part of the pupil is essential. One plan which has been proved to be quite effective involves the pupils' use of handwriting scales for the appraisal of their own writing. A copy of some good general merit scale should be conveniently placed in every classroom to encourage pupils in its use as a means of facilitating comparisons and evaluation of personal products.

Another means of motivation is the exemption from further penmanship drill of all pupils who have attained the accepted standard of speed and quality. The standard of 60 for speed and quality is the one generally accepted. Evi-

[41] Freeman, Frank N., *Experimental Education,* Houghton Mifflin Company, 1916.

dence seems to indicate that from 50 to 75% of the pupils in the upper grades can easily reach this standard. If these pupils are exempt from further drill, the teacher is able to devote more time to those who have failed to meet the standard.

Discussion Exercises

1. From the standpoint of classroom emphasis which should receive the most attention, oral or written language?
2. How does your answer conform to the actual facts in the case as shown by relative social demands and by actual instructional emphasis in the classroom?
3. What do you consider to be the major reasons why written English abilities receive most of the attention in the upper elementary and high school grades? Why do teachers avoid emphasis on the oral language skills?
4. In your opinion would it be practicable to prepare a series of oral language scales for use on phonographs in the classroom? Would they have any real value for testing? For teaching?
5. Which of the several tests described and discussed in this chapter sample most widely into the skills listed in the outline of skills on pages 313 to 315? Does this offer any guarantee of the validity of the tests?
6. In your judgment are there any diagnostic language tests? Why?
7. Follow the suggestions given in the final pages of this chapter and prepare sets of remedial exercises for at least two language skills other than the ones listed.
8. What appears to be the most acceptable fundamental assumption upon which the spelling vocabulary suitable for high school instruction should be based?
9. Why may words selected from standard spelling scales and similar sources not be used as the basis for measuring spelling ability in any given class? Under what conditions would their use be valid?
10. What range of difficulty in words would you select for the purpose of measuring a class of unusually poor spelling ability?
11. Prepare a set of dictation sentences using the words given in the list test on Page 336 and establish the time dictation rate for these words in an eighth grade class.

12. Make a list of all of the factors you are able to identify as operating to affect the quality of a pupil's handwriting.
13. From the evidence you have seen and the facts presented in this chapter is there a direct relationship between the quality of handwriting and the rate at which it is produced?
14. Would you expect to find wide differences in the abilities of different teachers to rate handwriting specimens for quality? Does this suggest anything to you regarding the need for this type of discriminative power on the part of the teacher?
15. Secure and rate at least a dozen samples of handwriting representing widely varying levels of quality. Record your scores on a separate sheet of paper. After a delay of two weeks score the same samples again. On what percentage of the samples did you agree with your previous marks within five points on the Ayres Scale?

Selected References

Ashbaugh, E. J., *Iowa Spelling Scales.* Bloomington, Illinois: Public School Publishing Company, 1919.

Ayres, L. P., *A Measuring Scale for Ability in Spelling.* New York: Russell Sage Foundation, 1915.

Ayres, L. P., *A Scale for Measuring the Handwriting of School Children.* New York: Russell Sage Foundation.

Ballenger, H. L., *The Validation of the Iowa Elementary Language Tests.* University of Iowa Studies in Education. Iowa City: University of Iowa, 1931.

Breed, F. S., *How to Teach Spelling.* Dansville, New York: F. A. Owen Publishing Company, 1930.

Brueckner, Leo J., and Melby, Ernest, *Diagnostic and Remedial Teaching,* Boston: Houghton Mifflin Company, 1931.

Burton, Richard, *Why Do You Talk Like That?* Indianapolis: The Bobbs-Merrill Company, 1929.

Cline, E. C., *Your Language.* New York: D. Appleton and Company, 1930.

Davis, Georgia, "Remedial Work in Spelling," *Elementary School Journal,* Vol. 17, March, 1917.

Freeman, F. N., *Chart for Diagnosing Faults in Handwriting.* Boston: Houghton Mifflin Company, 1914.

Greene, H. A., Betts, E. A., "A New Technique for the Study of Oral Language Activities," *Elementary School Journal,* Vol. 33; No. 10, June, 1933.

Greene, H. A., *A Criterion for the Course of Study in the Mechanics*

of Written Composition, University of Iowa Studies in Education, 1931.

Greene, H. A., "Research in Elementary Language," *Elementary English Review*, March, April, May, and June numbers, 1933.

Horn, Ernest, *A Basic Writing Vocabulary*. University of Iowa Monographs in Education, First Series, No. 4. Iowa City: University of Iowa, April, 1926.

Lyman, R. L., *Summary of Investigations Relating to Grammar, Language, and Composition*. Supplementary Educational Monographs. Chicago: The University of Chicago, January, 1929.

McKee, Paul, *Language in the Elementary School*. Boston: Houghton Mifflin Company, 1934.

National Education Association, *Fourth Yearbook of the Department of Superintendence*, pages 113-114, Washington, D. C., 1926.

Nystrom, Ellen, *The Self-Corrective Tests in Handwriting*. Minneapolis: Farnham Press.

Odell, C. W., *Educational Measurement in High School*. New York: The Century Company, 1930.

Smith, H. L., and Wright, W. W., *Tests and Measurements*. New York: Silver, Burdett and Company, 1928.

Stone, Clarence R., *Supervision of the Elementary School*. Boston: Houghton Mifflin Company, 1929.

Taylor, J. S., *The Supervision and Teaching of Handwriting*. Richmond: Johnson Publishing Company, 1926.

Thorndike, E. L., "Handwriting." *Teachers College Record*, New York: Columbia University, March, 1910.

Travis, Lee Edward, *Speech Pathology*. New York: D. Appleton and Company, 1931.

Ward, C. H., *Grammar for Composition*. Chicago: Scott, Foresman and Company, 1934.

West, P. V., *Changing Practices in the Teaching of Handwriting*. Bloomington, Illinois: Public School Publishing Company, 1927.

Wheat, H. G., *The Psychology of the Elementary School*. New York: Silver, Burdett and Company, 1931.

Wilson, G. M., and Hoke, K. J., *How to Measure*. New York: The Macmillan Company, 1928.

Wise, Marjorie, *On the Techniques of Manuscript Writing*. New York: Charles Scribner's Sons, 1924.

CHAPTER XIV

MEASUREMENT IN READING AND LITERATURE

THIS chapter is a summary of the following measurement and corrective problems involved in instruction in high school reading and literature:

a. Educational and social significance of reading in the high school.
b. Major objectives of reading instruction.
c. Essential reading skills.
d. Measurement of reading abilities.
e. Remedial and corrective work in reading.
f. Reading and literature.
g. Objective tests in literature.

I. EDUCATIONAL AND SOCIAL IMPORTANCE OF READING

181. **Reading as a Tool Skill.** The solution of most classroom problems in the modern school requires the skillful use of books as sources of information. In fact the ability to read is not only fundamental to progress in school but is essential to successful participation in the activities of life. Thus reading is something more than merely the rapid comprehension of printed symbols and the memorizing and organization of the materials read. It is the ability to make efficient use of books and libraries as sources of information. As Gray[1] has pointed out, reading is an indispensable means of "familiarizing adults with current events, with significant social issues, with community and national problems, and with American institutions, ideals and aspirations. It is also essential in attaining vocational efficiency, in broadening one's

[1] Gray, W. S., "Importance of Intelligent Silent Reading." *Elementary School Journal*, January, 1924.

range of information, and in seeking pleasure and profit during leisure hours." This tendency to treat reading as a most important tool of learning has resulted in establishing a very close relation between reading and practically every other school activity.

The necessity for a high level of reading ability on the part of high-school and college students is more readily appreciated when it is recognized that approximately ninety per cent of the bulk of facts these students are expected to master are obtained from books, or at least as a result of reading. It has too often been assumed that pupils have mastered all of the basic reading skills in the elementary school and that consequently the high school teacher of English may devote his reading emphasis exclusively to the technicalities, beauties, and subtleties of the classics and of modern literature. Numerous investigations have shown conclusively that many high-school and college students are gravely handicapped in reading abilities. Deficiency in reading ability is a potent cause of school failure, not alone in English but in practically all academic subjects. Dr. W. F. Book concludes that "it is clear why many college students fail to do satisfactory work in their college courses. Some are so deficient in their ability to get points made by the authors they read and from whom they must get the facts to be learned that we have little reason to expect success from them until this deficiency is successfully made up."

As a result of a program of special instruction given to 54 students under his direction Dr. Book[2] reports in this same article that the reading efficiency of these individuals was improved 102 per cent during a semester. Their ability to master their assignments was increased from 60 per cent to 97.3 per cent efficiency. Some individuals improved their reading efficiency as much as 250 per cent. Other recent data of a similar nature indicate that the reading abilities of adults can be markedly improved through systematic use of properly devised drill. Averill and Mueller[3] report that

[2] Book, W. F., "How Well College Students Can Read." *School and Society*, Vol. XXVI, August 20th, 1927; pages 242-248.

[3] Averill, L. A., and Mueller, A. D., "The Effect of Practice on the Improvement of Silent Reading in Adults." *Journal of Educational Research*, Vol. 17; No. 2; February, 1928; pp. 125-129.

sixteen adult subjects were able practically to double their rate of reading without decreasing their comprehension as a result of three forty-minute practice periods per week for approximately three months. These results compare closely with the findings of O'Brien in his experiments with Illinois school children.[4]

The results of such investigations serve to reveal at least two important facts. First, the abilities of mature students and adults to apply their reading skills effectively are not in keeping with our assumptions. Second, the situation is not hopeless, for the evidence shows, that, given proper diagnosis and the systematic application of properly constructed instructional and drill materials, they can develop a significant amount of improvement in reading ability in a relatively short time.

182. **Reading and Literature.** The use of reading for the acquisition of facts and figures, for the satisfaction of academic requirements, represents only one small phase of its function. The ability to read opens up rich avenues of enjoyment and pleasure in the fields of literature which provides the mediums through which the individual can become acquainted with life, its meaning and significance. It portrays those aspects of life, its activities, ideas, ideals, emotions about which human interests cluster. But these are not freed for the individual unless he has acquired skill in reading and a desire to sample the right sort of experiences in his reading. Thus the teacher of high school and college English faces the dual responsibility of developing new or reviving old reading skills, and directing the reading interests of the individual into wholesome experiences.

183. **Major Objectives in Reading Instruction.** The life situations under which reading is usually done may be conveniently grouped into two types, depending upon the attitude of the reader. These are:

1. Reading for pleasure, or reading of the leisure type.
2. Reading for information, or reading of the work-level type.

[4] O'Brien, J. A., *Reading, Its Psychology and Pedagogy.* The Century Company, New York, 1926.

It must be clear that no hard and fast lines can be drawn between these two levels of reading activity since an individual may shift from one type to the other without realizing the change. However, the distinction between the type of situations under which reading ordinarily takes place affords a convenient basis for the analysis and location of specific reading skills around which to develop diagnostic and remedial materials.

The accompanying outline of the field of reading is adapted from an outline of reading skills prepared by Horn and McBroom,[5] and supplemented from a functional analysis presented by Yoakam [6] on pages 54 to 56 of his "Reading and study," an excellent treatment of the problem of securing more effective study methods through the development of better reading habits. A number of the uses reported by Gray as a result of his analysis [7] of students' reading activities in the University of Chicago High School, are also utilized in this outline.

A FUNCTIONAL ANALYSIS OF READING

A. Reading of Leisure Types
- I. Oral Reading
  1. entertainment of others
  2. self-entertainment
  3. appreciation of beauty of expression
  4. practice on expression
- II. Silent Reading
  1. self-improvement
  2. self-entertainment
  3. vicarious experience
  4. appreciation
  5. practice

B. Work-Types of Reading (Major portion of reading time)
- I. Oral Reading
  1. entertainment of others
  2. instruction of others

[5] Horn, Ernest, and McBroom, Maude, "A Survey of the Course of Study in Reading," University of Iowa Extension Bulletin No. 99, February, 1924, College of Education Series No. 3.

[6] Yoakam, G. A., *Reading and Study*, The Macmillan Company, New York, 1928.

[7] Gray, W. S., "The Relation between Study and Reading," *Proceedings of the National Education Association*, Vol. LVII, 1919; pages 580-86.

3. acquisition of new modes of expression
4. extension of vocabulary

II. Silent Reading (Major aspects of work-types of reading)
1. Examples of specific situations in which one reads silently
a. to secure the facts on which judgments may be based
b. to secure information on the essential conditions of a problem to be solved
c. to verify a fact or an opinion
d. to secure a basis for action
e. to secure general information in desirable fields
f. to acquire information for specific use
g. to evaluate material as to its applicability to a specific case
h. to secure a basis for the formulation of an opinion from data or statements
i. to analyze the essential elements of an argument
j. to enlarge one's vocabulary
k. to acquire more effective modes of expression
l. to master technical and contextual meanings of words
m. to acquire a basis for a critical questioning attitude on a controversial issue
n. to provide a basis for the understanding of a specific situation
o. to satisfy the desire to discover new problems and answer them

III. Essential Skills Involved in Work-Study Types of Reading, and Suggested Methods of Acquiring Them

1. Skills and Abilities	2. Methods [8]
a. Skill in recognizing new words.	*a.* Methods suggested in manuals of standard method readers.
b. Ability to locate material quickly:	*b.* —
(1) Knowledge of and ability to use an index.	(1) Children learn alphabet, find words in alphabetical arrangement, find words in an index, find answers to questions by use of index, make an index for a book which has none.

[8] Quoted with permission from Horn-McBroom, *op. cit.*

(2) Ability to use a table of contents.	(2) Assign lessons by title; find lessons in table of contents; find authors in table of contents.
(3) Ability to use the dictionary.	(3) Handled in detail in Rice, O. S., *Lessons on the Use of Books and Libraries*, Rand, McNally and Company.
(4) Ability to use library card files.	(4) Handled in detail in Rice, O. S., *Lessons on the Use of Books and Libraries*, Rand, McNally and Company.
(5) Ability to use reference material.	(5) Handled in detail in Rice, O. S., *Lessons on the Use of Books and Libraries*, Rand, McNally and Company.
(6) Ability to use keys, tables, graphs, etc.	(6) Practice in interpreting maps, tables, etc., in geographies, etc.; make graphs or tables to illustrate problems.
(7) Ability to skim.	(7) Certain topic given, children skim to find material about it, to find answer to a question, a sentence which proves point, etc.
c. Ability to comprehend quickly what is read: (1) Rhythmic and rapid eye movements. (2) Absence of lip reading.	*c.* Use flash cards with words, phrases, sentences, requiring action responses; find answers to questions; answer true and false state-

(3) Knowledge of meaning.	ments; follow directions; completion sentences; labelling objects; matching words and objects, or words and pictures.
d. Ability to select and evaluate material needed.	*d.* Same as "c"; also practice noticing date of publications, name of author; choose from several statements one which answers question best.
e. Ability to organize what is read: (1) To summarize. (2) To assign topics to proper order or place. (3) To discover related material. (4) To outline.	*e.* Practice giving subjects to a paragraph; choose paragraph which answers questions; find everything which bears on a topic; put related words together; give a summary; outline a lesson or page.
f. Remembrance of material read.	*f.* Analysis of how to memorize quickly; practice in it; points in an article memorized; reports made without notes.
g. Knowledge of sources.	*g.* Practice using common sources: dictionary, encyclopedias, yearbooks, magazines, readers' guides, etc.; make lists of sources.
h. Attitude of attacking reading with vigor.	*h.* Hold every child responsible for jobs assigned; tie up reading with project work; show each child his needs through testing.
i. Attitude of proper care of books.	*i.* Lessons on how books are bound; what ruins bindings; how to keep books clean, how to make them endure; who pays for the books, etc.

184. Measurement of Work-Study Type of Reading. The emphasis given to the work types of reading in the skills listed on pages 352 to 354 indicates something of the importance of this type of reading in relation to the total reading field. Some pupils fail (in first year algebra, for example), not entirely because of ignorance of the basic facts, or lack of mental ability to understand the explanations, but rather on account of sheer inability to read. In fact, one of the best ways to improve work in many other school subjects is to make a drive on the work-type of reading ability. A recognition of this has caused recent makers of tests in reading to turn their attention in this direction. A number of excellent reading tests which provide useful analytical information are available for use in high school and college classes. Among the more widely used of these tests suitable for the junior and senior high school are the *Detroit Reading Tests* (Test 4); *Gates' Silent Reading Tests, Advanced*; *Haggerty Silent Reading Examinations, Sigma 3*; *Iowa Silent Reading Tests, Advanced*; *Sangren-Woody Reading Tests*; the *Stanford Achievement Reading Tests*; the *Unit Scales of Attainment in Reading*; the *Whipple High School and College Reading Test*, *Thorndike-McCall Silent Reading Scales*, etc. Of these the *Iowa Silent Reading Test, Advanced*, and the *Whipple High School and College Reading Test* are standardized for college use also.[9] Only a few samples of these tests will be described here as examples of procedures in the measurement of silent reading abilities.

The *Iowa Silent Reading Tests* (*Elementary* for Grades 4 to 9; *Advanced* for high school and college students) are among the most recent and comprehensive tests designed to provide a detailed and analytical measure of silent reading abilities. These tests go beyond the general survey of two or three phases of silent reading ability. They cover a wide sampling of the skills essential to effective reading of the work-study type. Naturally they do not succeed in measuring all of the major objectives of reading as outlined on pages 352 to 354 of this chapter. The specific skills which

[9] The publishers and distributors of tests mentioned in this volume will be found in pages 585-595 of the *Appendix*.

are measured by the tests are given on page 107 of this book. The tests sample into numerous reading skills and into different subject-matter fields. For example, *Test 1, Paragraph Meaning*, consists in two parts, one dealing with science content and the other with literature. *Test 2, Word Meaning* is made up of four parts sampling systematically into the special subject-matter vocabulary of social science, science, mathematics, and English, respectively.

One of the more valuable reading skills measured by these tests is the ability to use the index to locate information. This test, which is Part B of Test 5, is reproduced here in order to suggest how this ability may be measured. It should also give to many teachers hints for the development of similar material for instructional purposes. The advanced examination of the *Iowa Silent Reading Tests* is standardized for grades 9 to 12 and college freshmen. The elementary examination is standardized on a grade and a mental-age basis for grades 4 to 9 of typical eight year elementary school systems and for grades 4 to 8 of seven-year elementary systems. The testing time for the elementary test is 42 minutes and for the advanced test is 35 minutes. Reliability coefficients of .90 and .95 are reported for the Elementary Examination and the Advanced Examination respectively.

Test 5. Location of Information

PART A. USE OF THE INDEX

Directions. This is a selection of items from an index. It is needed in connection with the exercises which follow, on the use of the index.

Index

Alaska: agricultural possibilities, 213, 214; commerce, 214, 215; exports, 214 (Fig. 147); fisheries and forests, 210-212; fur farms, 210; imports, 214; map (Fig. 129), facing 197.

Communities: definition, *see* Community; kinds of, 9-18; origin of, 11-12; growth of, 12-20; large and small, 15-18; co-operation of, 19-20, 42, 49, 113; and health, 31-43; and the protection of life and property, 45; and education, 65-66; and beauty, 72-85; money for expenses, 87-98; organization of, 99-111; dependent upon each other, 139.

Corn: in America, 87; Argentina, 282; United States exports (Fig. 190), 282-284; plant, 87; industrial uses of, 189.
Courts: city, 108-110; county, 125; state, 120-127; national, 260-268.

Dairy products: 82-84; in northern states, 82-83; states leading in production, 83. *See also* Cattle.
Denmark: commerce, 411, 412; exports, 80, 411 (Fig. 262), 411; map (Fig. 258), 408; possessions, 242; resources, 81-84.
Dyes: 148-155; ancient, 148-150; aniline, 153-155; cochineal, 150-152; discovery of native American, 150; dyewoods (*see* Forest); manufacture in the United States, 155; recent increase in value (Chart 15), 154; substitutes for natural, 153, 226; value of aniline, 154-155 (Fig. 20), 158. *See also* Cochineal.

Flour. *See* Wheat.

Nervousness: 85; and school life, 108; physical defects, 110; preventable, 111; causes of, 112; habit, 111, 113; from tobacco, 363.

Wheat: fertilizers for, 48; insects injurious to, 51; kinds of, 43; preparation of seed bed, 45; when to sow, 47; world's bread grain, 42; wheat diseases, 49.

Study the following samples:
Where (on what page as shown in the index above) will you find information on coal mining in Indiana? . . . (145)
Can you find information regarding oil in Indiana? . (Yes)
Answer the remaining exercises in a similar manner.

---

1. Next to what page can you find a map of Alaska? . . ( ) 1
2. Does the index tell where to find information about the industrial uses of corn? . . . . ( ) 2
3. What is the number of the figure which shows something about the export of corn by the United States? . . . . . ( ) 3
8. Does the index tell you on what page you can find something about flour? . . . . . . ( ) 8
9. On what page can you find a figure about the exports of Denmark? . . . . ( ) 9
10. On what pages is the subject of dyes discussed? . . . ( ) 10

Do not work on the next part until told to do so.

Number right . . . . . × 2 . . . . (*Score, Test* 5 *A*)

The *Sangren-Woody Silent Reading Test* is one of the better analytical types of reading tests designed for use in the upper elementary grades. The test measures seven specific aspects of silent reading ability; namely, word meaning, rate, fact material, total meaning, central thought, following directions, and organization. Illustrations of the types of exercises used in two of the parts are given below:

PART I. WORD MEANING[10]

(Forty Exercises ; Time limit — three minutes)

1. A dress is something to 1 drink 2 eat 3 wear 4 throw ( )
40. Chronicle is a kind of 1 disease 2 poetry 3 king 4 record ( )

. .

PART III. FACT MATERIAL

(Twenty questions ; Time limit — three minutes)

Directions : Write the answer to each question on the dotted line. Use one word if possible.

The "lead" in your pencil is not made of lead. Long ago people had lead in their pencils ; that is probably why the pencils we use are called lead pencils. Another mineral called "graphite" is now used. This mineral is taken from mines in the same way as coal or iron ore.

1. What did people use in their pencils long ago ?
2. What mineral is used in the pencils now ?
3. From what is the mineral graphite taken ?

Whipple's High School and College Reading Test is designed to measure only a single factor of reading ability, the comprehension of relatively difficult content material. The test utilizes material suited to mature readers and demands continuous and relatively long attention to a single topic. Comprehension questions are scattered through the context in a way to insure careful reading. The performance of the reader is expressed in a single score earned within the twenty minute time limit set for the test.

The *Nelson-Denny Reading Tests for Colleges and Senior High Schools* includes a one-hundred word vocabulary test based upon the student's reaction to multiple-response exercises, and a comprehension test made up of nine paragraphs chosen from the fields of literature, history, science, and social science. The working time is thirty minutes. The reliability of the test is .914 based on the reactions of college freshmen. The tests differ from most tests in that they are provided with a self-marking device.

The *Poley Precis Test* is designed to measure the student's ability to indicate which of several summaries or abstracts of paragraphs most nearly expresses the main thought

[10] Reproduced with the permission of the publishers, World Book Company, Yonkers-on-Hudson, New York, 1927.

of the selection. It is available in two forms, each form containing eight selections. Each selection is followed by five summaries or precis expressing in different degrees of quality the main thought in the paragraph. The selections used in the test exemplify widely different fields of interest. The instructor will find the tests useful as teaching and practice exercises also. The student can thus be brought to see that the English classroom embraces all fields of human interest, and that the ability to grasp the central idea of a passage in science or history is as important as it is in literature or general reading. The tests and the subsequent discussion of results should also provide a motive for further writing of precis in the fields which are most difficult for the student. The reliability of the test is not high although the working period is forty minutes.

185. Remedial Drills for Work-Study Types of Reading Difficulties. Possibly one reason for the rather marked instructional emphasis on the work-study type of reading to the exclusion of reading of the leisure types is that there is good reason to expect a considerable carry-over of skills from work-type reading to the other type, because of this large number of common skills involved. For example, skill in recognition of word meaning which functions in work-study reading is probably similarly effective when the individual is reading solely for pleasure.

A few illustrations of specific types of remedial exercises suited for use in silent reading of the work-study types are presented in the following pages in the hope that they may serve as a guide to teachers interested in the development of material of this type. Only a few samples from each field can be furnished.

1. Practice similar to the procedure used in Test 1 of the *Iowa High School Silent Reading Tests* will aid the teacher in directing the pupil's attention to the need for greater care in the reading of paragraph units. Almost any supplementary booklet or bulletin may be readily prepared for this purpose. An examination of Parts A, B, or C of Test 1 of the tests will make clear how this material may be made. The questions may be mimeographed on separate sheets or

may be written on the blackboard. This type of material is extremely useful as supplementary drill during the supervised study period.

2. Teachers should assume responsibility for teaching the use of keys, tables, maps, graphs, etc. The student of science, social science, or mathematics, can scarcely utilize all the available information until he has acquired skill in the interpretation of these devices. Graphic representations are becoming increasingly popular for the presentation of certain types of factual data. Newspapers and magazines continually use them. The student should be given extensive practice in the construction and presentation of this type of reading material.

General Suggestions on Comprehension

1. Hold the pupil to definite tasks. This means that well-planned and definite assignments must be made.
2. Every class affords a splendid opportunity for the formation of the proper attitude toward books. Discuss with the pupils the use, care, and desirable mechanical features of a book.
3. All subjects are interrelated. All high school teachers can motivate the acquisition of language and reading skills by the assignment of written problems which are of basic interest in their particular courses.
4. Let the pupils know their specific weaknesses, and how you propose to help them.
5. Any change in the reading program must take place largely in the assignments and recitations of the content subjects.
6. Whipple's *How to Study Effectively*[11] is a valuable basic book for Social Science and English classes.

Suggested Remedial Procedures for Organization Difficulties

1. Explain a standard outline form if necessary.
2. Give practice in outlining paragraphs, sections, chapters, and other mechanical units of the subject matter. Make use of the type of drill illustrated in Test 5, Part B of the *Iowa High School Silent Reading Tests.*
3. Give short quizzes which require the summarization of some unit of work. This may be over a part or all of the day's preparation or some other unit of time or subject matter.

[11] Whipple, G. M., *How to Study Effectively,* Public School Publishing Company, Bloomington, Illinois, 1916.

4. Create an interest in debate which requires the selection and presentation of facts about some single idea.
5. Assignment topics which can be answered after the pupil has skimmed the materials should be given. This is a reading skill used in many life situations.
6. The mathematics teacher may develop a specific skill when a series of facts is required to be organized in order to reach the solution.
7. Give practice in reading a selection to get the main points clearly in mind. When this particular type of exercise is used give no questions or problems as guides. Impress upon them the necessity for making a personal evaluation and selection of the important points.
8. Give drill in compiling selected bibliographies from other bibliographies. This gives important training which will aid the child in himself suggesting reliable sources of information.
9. Give drills in answering the question "What does the author attempt to answer in this lesson?" This type of exercise may be carried out relative to the book, a chapter, a section, or a single paragraph.
10. Almost any of the content subjects lend themselves to practice in the selection of the central idea of a paragraph. A teacher can construct practice materials for this type of reading with a minimum of effort.

186. **Other Remedial Materials in Reading.** The recent analytical work of Gates, Gray, Betts, Durrell and others opens up great possibilities for diagnostic and corrective work. No longer is diagnosis in reading obliged to confine itself to such vague and general qualities as rate, comprehension of word meanings, etc. In fact, it now becomes quite apparent that the vast majority of so-called diagnostic tests in this field are not at all suited for the specific types of diagnosis required in the identification of reading disabilities. Excellent commercial materials designed for instructional testing purposes and remedial uses in reading are available in many different forms. At the present time there are almost countless drill books and work-books which aim to develop silent reading skills of the work-study type.

187. **Vocabulary Tests.** One of the major aims of instruction in all phases of English is the development of a rich and expressive vocabulary. While vocabulary tests are not necessarily classified as reading tests they are related to them in nature and function. A few vocabulary tests are summarized here as suggestions to the English teacher.

The *Markham English Vocabulary Test*[12] is available in two forms, Form 1 and Form 2. It is designed for use in high schools and colleges. The 125 test words comprising each form of these tests represent a sampling of the field covered by the general reading vocabulary of both high school and college students. The test words were selected from a reservoir of about 1,000 words obtained by taking the fifth word on each even-numbered page and the last words of each odd-numbered page of the *Winston Simplified Dictionary* and by adding seventy-five words taken at random from Webster's small dictionary and a few words selected at random from students' manuscripts. The words used in the test should serve as a basis for an effective program both of teaching and of testing. The coefficient of self-correlation for Form 1 and Form 2 is .88, P.E. ±.0047. The test can be completed in one class period.

The *Holley Sentence Vocabulary Scale, Series 3B* is composed of seventy sentences. The pupil underscores one of four words which he considers as correctly defining the related word in the sentence. The scale is intended for use in grades 7 to 12. The reliability is reported as .82 on 140 high school students, all grades, with a probable error of score of ± 2.4.

The *Pressey Technical Vocabulary* lists have proved their value in determining the vocabulary abilities of pupils in the different subject fields. A word list is available for each of the following fields:

1. Grammar and composition (English, Latin, French, German)
2. Literature
3. Arithmetic
4. Algebra and geometry
5. American history
6. General science
7. Biology
8. Chemistry
9. Geography

These lists were prepared in accordance with teachers' ratings of the importance of technical words appearing in common text books. The author of these lists suggests that

[12] All tests mentioned in this volume are listed with publishers in the *Appendix*.

"each pupil go through the list for the subject being taught and systematically master every word." The author further states, that "these lists are more than mere lists of words; they are catalogues of the important concepts in each subject. So the lists may very profitably be used as diagnostic tests in determining the specific weaknesses of a class or an individual pupil in a subject. That is, if the teacher will find out the terms which the pupil does not know, in the vocabulary for the subject in question, she will have a very valuable indication as to the topics on which the pupil needs help. The lists can be of great service in thus diagnosing weaknesses, and guiding the teacher in her remedial instruction."

The *Inglis Test of English Vocabulary* is available in two forms, A and B. It is intended for use in grades 9 to 12 and in college classes. No time limits are set. The test contains 150 sentences in each of which one word is in italics. In the list of words after each sentence the pupil is asked to underline the one word that has the same meaning as the italicized word.

II. Measurement in Literature

188. Objective Tests in Literature. The measurement of outcomes of instruction in literature has been confined almost entirely to the informational side, since the more mechanical features of instruction are the ones most readily measured by tests, but recently something has been accomplished in the measurement of the appreciative and aesthetic values of literature. Descriptions of a few representative tests in this field are given in this section.

The *Analytical Scales of Attainment in Literature* provide separate measures of information about authors, literary characters, and the general field of literature. They give the teacher a comparison of the attainments of students in each of these phases of instruction in literature and show the phases on which further emphasis needs to be placed. The time limit for these scales is forty minutes.

The *Objective Tests in English* prepared by Mabel S. Satterfield and a group of nine collaborating authors work-

ing under the editorial direction of Dr. George A. Rice cover the American and English classics most frequently studied as a college entrance requirement. Tests covering fifty-four titles are now available. The tests are designed to reveal the ability of the student to judge facts, analyze and form conclusions following a study of these classics. The majority of the tests include approximately one hundred items sampling into such fields as information concerning the authors, the vocabulary of the classic, the recognition of quotations, the interpretation of characters, and appreciation of background. No time limits are specified by the authors but it is apparently assumed that for most students a test will occupy a typical class period.

The *Logasa-Wright Tests for Appreciation of Literature* are designed for use in high school and college. The series is composed of six tests: (1) Discovery of Theme, (2) Reader Participation, (3) Reaction to Sensory Images, (4) Comparisons, (5) Trite and Fresh Expressions, and (6) Rhythm. The tests are arranged in the order in which the student learns to appreciate literature. Naturally, he begins with the discovery of theme, since a selection means nothing to him until he knows what it is all about. From that he goes to reader participation, although he may not signify his reaction to the selection by any such name. In turn he comes to the elements which lead to that participation, viz., reaction to sensory images, a discrimination between good and poor comparisons, an appreciation of freshness in expression as opposed to triteness, and a recognition of rhythm in the production.

The two forms of the *Abbott-Trabue Scales for Judging Poetry* each comprise thirteen sets of stanzas. Each set includes in addition to the original poetic arrangement, three derived stanzas. The derived ones represent varying degrees of mutilation of emotional quality, imagery, and meter in the original. The pupil is asked to read the four poems in each set and to decide upon the one he likes best and the one he likes least. The test is quite short and its reliability suffers from this fact as well as from the method used in scoring the test.

The *Carroll Prose Appreciation Tests* are standardized

for use in Junior and Senior High Schools. They measure one of the great outcomes of instruction in literature, the ability to appreciate prose. They do so by measuring the pupil's ability to differentiate good prose from poor, and poor from very bad. These tests are made up of a series of sets, each set containing four selections taken from the following sources: First choices, from books generally regarded as of excellent quality; second choices, from books generally regarded as of poor quality; third choices, from the less literary magazines; fourth choices, from mutilations. The task of the pupil is to arrange these excerpts in their proper order. The degree to which he accomplishes this indicates quantitatively his ability to appreciate prose literature. Percentile norms established on 4,200 cases are available for junior and senior high school grades.

The test sheets are usable repeatedly. Only record blanks need be reordered. The testing time is approximately 35 minutes although the test is not a timed test.

189. **Reading Scales in Literature.** Many students fail to do successful work in literature because of their inability to read and comprehend the ideas in the lessons assigned. The following tests which are briefly described are designed to enable the teacher to determine how well the student can grasp the content of literary material.

The *Van Wagenen Reading Scales*[13] are available in the following forms: *A*, *B*, and *C* for testing the pupil's ability to comprehend what is read; and *Alpha* and *Beta* for determining the pupil's ability to interpret what he reads. Each test consists of fifteen paragraphs related to the field of literature. After each paragraph are several statements concerning the context. The pupil is asked to check the true statements. These scales are intended for use in grades 8 to 12 inclusive.

The *Stanford Test of Comprehension in Literature* is available in two forms, Forms A and B; and Tests 1, 2, and 3. Each test consists of ten paragraphs considered of the average difficulty of the books (Standard Classics) from which each is taken. After each paragraph, four multiple choice questions calling for a grasp of the thought and spirit

[13] *Ibid.*

of the passage are included. The test is similar in purpose to the *Van Wagenen Reading Scales* in that it is intended to determine the particular reading skill needed for the effective study of literature.

Discussion Exercises

1. In what ways does modern life place a particular burden on the ability to read rapidly and well?
2. What are the essential differences between reading for appreciation as in literature and reading for acquisition of facts as in the study of the content subjects?
3. In what high school subjects would you expect achievement to be definitely affected by reading disabilities? Why?
4. Check the list of Essential Skills Involved in Work-Study Type Reading against the skills specified for measurement in the Sangren-Woody, the Haggerty Sigma 3, and the Iowa Silent Reading Tests.
5. What emphasis should oral reading be given in the high school? Where does it belong, in English or in speech training?
6. Study the index test shown on page 357 and prepare a set of suggestive drill exercises for the development of skill of this type.
7. From your reading and observation make a list of the major reading defects as encountered among high school students.
8. Prepare sets of remedial exercises suitable for high school use following out two of the suggestions for drills on organization difficulties (see page 361).
9. To what extent is literary appreciation due to information? Do we like the things we know about, and therefore appreciate the things about which we are well-informed?
10. Prepare a list of specific comments evaluating available reading and literature tests.

Selected References

Betts, Emmett A., *Diagnosis of Reading Disabilities.* Teachers Guide for use with telebinocular equipment. Meadville, Pa.: Keystone View Company, 1934.

Book, W. F., *Learning How to Study and Work Effectively.* Boston: Ginn and Company, 1926.

Brueckner, L. J., and Melby, Ernest, *Diagnostic and Remedial Teaching.* Boston: Houghton Mifflin Company, 1931.

Buswell, G. T., "An Experimental Study of the Eye-Voice Span in Reading." Supplementary Educational Monographs, No. 17, December, 1920, University of Chicago Press.

Buswell, G. T., "Fundamental Reading Habits; A Study of Their Development." Supplementary Educational Monographs, No. 21, 1922, University of Chicago Press.

Durrell, Donald D., *The Diagnosis of Reading Disabilities.* Boston : Houghton Mifflin Company, 1935.

Gates, A. I., *The Improvement of Reading, Revised Edition.* New York : The Macmillan Company, 1934.

Gray, William S., and Monroe, Ruth, *The Reading Interests and Habits of Adults.* New York : The Macmillan Company, 1929.

Gray, W. S., "Remedial Cases in Reading ; Their Diagnosis and Treatment." Supplementary Educational Monograph, No. 22, University of Chicago Press, 1922.

La Brant, L. A., *The Teaching of Literature in the Secondary School.* New York : Harcourt Brace and Company, 1931.

McKee, Paul, *Reading and Literature in the Elementary School.* Boston : Houghton Mifflin Company, 1934.

Sangren, Paul V., *The Measurement of Achievement in Silent Reading.* Kalamazoo, Michigan: Bulletin of Western State Teachers College, 1927.

Smith, H. L., Wright, W. W., *Tests and Measurements.* New York : Silver, Burdett and Company, 1928.

Yoakam, G. A., *Reading and Study.* New York : The Macmillan Company, 1928.

CHAPTER XV

MEASUREMENT IN THE SOCIAL SCIENCES

THIS chapter summarizes the following points in the improvement and measurement of instruction in the social sciences:

a. Shifting objectives in the social sciences.
b. Measurable qualities in history, civics, and geography.
c. Measurement of acquisitive skills.
d. Measurement of outcomes of instruction.
e. Factual tests in history, civics, and geography.
f. Problem-solving and attitude tests in the social sciences.
g. Types of tests in the social sciences.

I. IMPORTANCE OF THE SOCIAL SCIENCES

190. **Objectives of Instruction in the Social Studies.** The subject matter of the social sciences deals primarily with the interactions of human beings as they associate with each other in varied political, economic, and social activities. Accordingly, the formulation of exact objectives in the social sciences is difficult. In fact, there are no scientifically established objectives, such as are found in many other school subjects. The relation of this fact to the problem of measurement in the social sciences is clearly shown in a statement by Horn:[1] "The most important single consideration in the whole problem of measurement in the social studies is the meaningful and authoritative description of the immediate objectives of instruction and of the specific instructional materials which are assumed to contribute to these and to the ultimate objectives."

The major objectives of the social studies as they have

[1] Horn, Ernest, "Another Chapter on Tests for the Volume on *Conclusions and Recommendations*." The Social Studies, January, 1935.

been stated in the past may be summarized in general terms as follows:

1. To develop proper social attitudes and conduct.
2. To inculcate an understanding of social environment, its character, needs, and problems.
3. To develop intelligent, unselfish, right-minded citizens.

It must be admitted that these objectives are extremely general and that they assume a liberal transfer from subject-matter content to actual behaviour which probably does not take place in actual life. According to Phillips,[2] "The social studies, which might be called the science of getting along with people, include the whole range of human activities, past, present, and future, with special emphasis on those facts and conditions which are concerned with men's well-being as they strive for the best which may be secured." Educationally, the ultimate objectives of instruction in the social sciences are:

1. To create ability and stimulate a willingness on the part of the individual to think in this field.
2. To provide a basis for the individual to draw useful and helpful lessons for the present and future from the past.
3. To develop a recognition and understanding of the origin and functions of present institutions, relationships, movements, and trends.

The difficulties involved in the effective measurement of these outcomes are obvious. The desired knowledges, skills, habits, ideals, and attitudes have not been sufficiently well identified to permit the accurate evaluation of achievement or progress.

The question of whether to organize the content of the social studies according to the traditional subject divisions or to integrate the various fields into a unified course of study is receiving much attention today. Unified courses are based on the theory that life consists of actual problems and that the best way to prepare children to meet such problems in life is to disregard subject divisions and assemble materials from all sources possible, putting it together in such ways

[2] Phillips, C. A., *Modern Methods and the Elementary Curriculum.* New York: The Century Company, 1931.

as may prove most effective in meeting the pupil's needs for problem solving. For the secondary school, this would mean ignoring economics, sociology, geography, ethics, political science, history and civics as separate subjects, and embodying related material from all of these subjects into an integrated course.

While there is considerable evidence favorable to the unification of the social studies, curriculum makers at the secondary school level have not made much progress in the development of such a course. Thus far these efforts have been sporadic and experimental. In the meantime, the schools generally will continue to teach subject-matter courses, with the result that subject-matter tests will probably still represent the best available material in these fields. Test construction in the integrated social sciences necessarily lags behind the development of the curriculum. As curriculum workers make progress with this idea, new tests will be developed in accordance with the objectives set forth in the best available integrated courses of study.

II. Possibilities of Measurement in the Social Studies

191. **Acquisitive Skills.**[3] Efficiency in the study of the social sciences depends to a marked degree upon the student's mastery of reading skills of the work-study type. Instead of providing only a few textbooks relating to a limited number of topics, the modern school provides wide reading opportunities as a means of enriching the course of study. Effective reading habits are therefore essential to such content subjects as history, geography, and civics. It follows, therefore, that good teaching in the social sciences, and in all of the content subjects for that matter, must provide for the improvement and refinement of the reading habits and skills required in these subjects.

Standard tests designed to measure a pupil's reading rate and comprehension in social-science content are not generally available. However, the teacher may develop his own informal tests of social science reading skill by formulating

[3] See *Glossary* for definitions of new terms.

comprehension questions based upon the reading material sampled from the pupil's textbook. A careful analysis of the pupil's answers and his rate of reading should be of great value to the teacher in improving the pupil's specific reading skills in the social sciences. The reader is referred to Chapter XIV for further suggestions in this connection.

The element of subject-matter vocabulary is basic to effective social science reading. Reading vocabulary, until recently, has not received the attention it deserves as a factor conditioning the efficiency of reading and study. Pressey[4] found that much of the difficulty which pupils have in studying their textbooks was due to lack of knowledge of the more or less technical words in a subject, rather than to lack of any general "silent reading ability." Tormey[5] has also shown that vagueness in the meanings which children attach to apparently simple phrases in history content definitely limits their accomplishment in the subject. He further showed that relatively brief periods of training in the acquisition of clear-cut meanings for these terms pay big dividends in greatly improved accomplishment in the subject.

192. **Measurement of Outcomes.** For most instructional purposes and for practically all testing purposes the basic knowledges involve the retention of a few facts with accuracy, others in approximation. These facts may deal with dates, men, events, or movements. To this series of facts may perhaps be added a rough idea of their time sequence and their relational interdependence. In civics the basic knowledge includes an understanding of the principles of government and their operation in the lives of men. Basic knowledge in geography includes memory of facts, skill in map interpretation, technical vocabulary, the interpretation of many and varied items of information in terms of their relationships, and the ability to solve problems dealing with human relationships.

The real deficiency in existing tests in the social sciences is not that they are designed primarily to measure the infor-

[4] Pressey, L. C., "The Special Vocabularies of Public School Subjects." *Educational Research Bulletin of the Ohio State University*, Vol. 3, April, 1924, Columbus.

[5] Tormey, T. J., "The Effect of Drill Upon the Specific and General Comprehension of Historical Content." Abstract in *Doctoral Theses in Education I*, Vol. IX, No. 1, pages 153-182, University of Iowa Studies in Education, Iowa City, January 15, 1934.

mational aspects of the subject, but that other abilities more important to the attainment of the larger objectives of social science instruction have not been provided for. Unfortunately, when a standardized test is used the particular outcomes which it measures tend to be given special attention by both teacher and pupils. As a result, important objectives other than those emphasized in the tests are likely to be neglected. The teacher must assume the responsibility for providing instruction on other types of outcomes. The measurement of such instruction will for the most part be necessarily taken care of by teacher-made tests.

The available body of facts in geography, history, and civics is large and is increasing rapidly. The difficulty of selecting the total array of facts deserving special emphasis has resulted in the production of relatively few tests in the social studies. In their efforts to meet the problem of which facts to teach and test, most workers in these fields have made their courses of study and their tests more and more comprehensive, hoping thereby to satisfy the ideas of all as to the basic items. Too often this has led both teacher and pupils to emphasize mere memorization of facts. As a result, these facts are too frequently mastered merely as facts and not because they may give the pupil a better understanding of life and human relationships.

193. **Kinds of Tests in History, Civics, and Geography.** The majority of the tests available at present in the social science fields are of doubtful value for diagnostic purposes. Three general groups of tests in the social studies may be identified:

1. Tests of facts and information.
2. Tests of ability to solve social problems.
3. Tests of civic, social, and economic attitudes.

Factual Tests. Tests of facts and information are by far the most numerous of the tests in the social studies, because the pupils' knowledge of such material is quite easily discovered, and because teachers of the social sciences, perhaps more than most other groups, have emphasized the acquisition of facts, and information, to the exclusion of other desirable

general outcomes of instruction. For diagnostic purposes, however, factual tests are of limited value. They fail to reveal why pupils do not know the facts if they had not been acquired. The factual tests aid the teacher very little in discovering the ability of students to use facts in their thinking in the social science fields. On the other hand, it must be admitted that there can be little thinking or problem-solving in any field without an adequate control over facts. No matter what the objectives may be, the knowledge of a certain body of facts is apparently fundamental to an understanding of the real purposes of instruction in the social sciences.

As an illustration of the measurement resulting from tests of this factual type, the accompanying sampling of items from the *Iowa General Information Test in American History*[6] is reproduced (Table 49).

Problem-Solving or Thought Tests. The development of the ability to utilize facts and basic principles in the attack on a novel social situation is one of the basic outcomes of teaching in the social studies. This type of problem-solving duplicates the steps in the ordinary process of thought. As in the case of arithmetic, problem-solving in the social studies involves reading the problem to comprehend it, picking out the facts which are pertinent to the problem, choosing a method of solution, and testing the results for accuracy and probability. Needless to say, this type of test in the social studies is very difficult to develop and most of the thought tests or problem-solving tests in these subjects leave much to be desired.

One of the earlier and more comprehensive attempts to analyze social science subject matter into its thought-provoking aspects is seen in the case of the *Barr Diagnostic Tests in American History.* These tests attempt to measure elements other than mere facts. They consist of five tests designed to measure (1) comprehension, (2) chronological judgment, (3) historical evidence, (4) evaluation of facts, and (5) causal relationships. The tests are designed for use in grades eight to twelve.

[6] All tests mentioned in this volume are listed with publishers in the *Appendix*, pages 585 to 595.

TABLE 49

CLASSIFICATION OF ITEMS IN *Iowa General Information Test in American History* ACCORDING TO TYPE OF ANSWER REQUIRED

Items Answered by:	Number of Items of Each Classification	
	Form A	Form B
Names of Men	18	17
Dates	5	6
Countries	5	10
Battles	2	3
Legislative Acts	11	8
Peoples	3	3
Wars	1	1
Industry	2	2
Names of Cities, Places	12	14
Names of Ships	6	2
Names of Groups or Parties	6	7
Attitudes or Ideas	1	4
Abstract Concepts	21	13
Numbers	3	3
Name of Special Thing	4	7
Total	100	100

The following is Item 1 of Test II, Chronological Judgment.

Indicate the chronological order (time order) in which the following men lived by placing a figure one (1) before the individual who lived first, a figure two (2) before the individual who lived second, and so on through the list.

(a) Samuel Adams
(b) U. S. Grant
(c) Sir Walter Raleigh
(d) Andrew Jackson
(e) William Penn

Attitude Tests. Since the actions of individuals depend to such a large degree upon attitudes and emotional reactions, the measurement of attitudes resulting from instruction in the social sciences is as greatly needed as are tests of ability to solve social problems. As a matter of fact, much attention is now being given in school to the development of desirable traits of citizenship which are so much needed in later adult life. However, it must be admitted that up to date the measurement of such traits by attitude tests has been a hope rather than a realization.

Part III of the *Brown-Woody Civics Test* deals with certain phases of civic thinking which call into play the expression of attitudes. Item 2 of Part III of this test is reproduced here as an illustration of this type of testing procedure.

BROWN-WOODY CIVICS TEST [7]

"Some of the members of your community are interested in securing a park for the use of the children in the west end of the city. You live on the east side of the city and would profit little, if at all, if the proposition should carry at the election. You are fully aware that a large sum of money is needed to purchase the property and that you must assume your share of the financial burden by paying a higher rate of tax on the property which you own. Why should you vote for the proposition?

1. Your property will increase in value.
2. Friends will benefit if the issue carried.
3. You may want to use the park.
4. It will entail but slightly higher taxes.
5. The park will benefit the community."

III. TYPES OF TESTS IN THE SOCIAL SCIENCES

194. Measurement of Achievement in American History.[8] In addition to the tests previously mentioned in this chapter a few of the more widely used tests in this field are described briefly. This is in no sense an exhaustive list of social science tests although many of the better tests in history, civics, geography and economics are listed.

The *Bowman United States History Test* is designed for use in grade seven and through the high school. Throughout the test the emphasis is placed on relationships. Only

[7] Reproduced with the special permission of the publishers, World Book Company.

[8] All tests described in this and other chapters in this volume are listed with publishers on pages 585-595 in the *Appendix*.

those personages, geographical items, dates, events and concepts which bear some specific relationship to the development of America are included in the test. The list of more than eight hundred items which constituted the basis for this test was compiled from seven or more objective lists and sources.

The *Columbia Research Bureau American History Test* covers the various aspects of American history approved by authorities for inclusion in courses for senior high school students. It is broad in its scope, including a sampling of 200 important facts, relations, and judgments in the several different phases of history. It is more than a memory test, being designed to measure ability to make judgments and to draw inferences from facts. Norms are given for high school seniors and college freshmen.

The *Denny-Nelson American History Test* is in two parts. Part I covers the period of discovery, exploration, and colonization ending in 1789. Part II covers the national period of American History from 1789 to present. Varying types of testing techniques are used, such as alternate answer questions, matching exercises, including the matching of cause and effect items and important dates and events, and completion items. Norms are given for each part separately for grades seven and eight.

There are two equivalent forms, A and B. The total working time for the test is 80 minutes. The correlation of the two forms is .92 for Part I and .93 for Part II.

The *Ely-King History Tests* comprise two tests in the series, (1) the *Ely-King American History Test* and (2) the *Ely-King Interpretation Test in American History.* The former test is intended for grades seven and eight. There is but one form composed of seven parts, each of which can be used separately after the units have been studied, or as a test of the entire year's work. Norms are available in terms of mental age and in percentiles. This test, which is largely informational, is valuable for both individual and group diagnosis. The *Interpretations Test* consists of two forms intended to measure the pupil's ability to interpret meanings and to evaluate historical situations, as well as reveal pupil-attitudes and appreciations. This test is provided with

percentile norms and is designed for use in the eighth grade.

The *Gregory Tests in American History* consist of two forms adaptable to grades seven and eight. In general the periods covered and the details called for are in line with current requirements in courses of study. These tests, while carefully prepared, fall short in certain of the large essentials which should characterize the outcomes of good history teaching. For example, some of the dates selected and the incidental facts called for in the test require memory of details that appear to be largely incidental to the main movements of history.

The *Harlan Test of Information in American History* is a brief informational test of the completion type. It contains ten exercises, calling for a knowledge of personages, historic terms, events, places, hypothetical conclusions, person-periods, dates-events, events-dates, statements and conclusions, significance of events, results of events.

The *Junior American History Test* covers the subject of American History as usually taught in the junior high school grades. It is broad in scope, including a wide sampling of important facts, relations, and judgments in the several different phases of history. It measures the student's ability to reason from facts as well as his accuracy in acquiring information. Norms and percentile ranks are given.

There are two equivalent forms, A and B. The time for giving the test is 60 minutes. The reliability coefficient is .92 for Form A and .81 for Form B.

The *Odell American History Scales* represent one approach to objectivity of measurement in the social sciences. Odell has developed a series of scales composed of pupil responses of nine types of thought questions: (1) analysis; (2) cause and effect; (3) comparison; (4) criticism; (5) discussion; (6) explanations; (7) relationship; (8) reorganizations; (9) summary. These scales of answers are in composition form and are taken from actual pupil responses to specific questions. The responses have been rated from 0 to 10. Following each answer in the scale a criticism is presented, showing its strengths and weaknesses. This scale should prove helpful to teachers in evaluating the answers of pupils to thought questions of such character.

The *Pressey-Richards Test on Understanding of American History* is composed of four tests: (1) character judgment; (2) historical vocabulary; (3) sequence of events; (4) cause and effect relationship. These tests are of the thought type, all prepared in one form, and may be used for grades six to twelve.

In Test 1 the objective is to measure the extent to which pupils have correctly appraised historical characters. Test 2 on Historical Vocabulary suggests a form of diagnosis of some significance.

The *Van Wagenen American History Scales* are eleven in number, including Thought Scale R2 in General History; Information Scales S1 and S2 in General History, and their alternate forms, R1 and R2; Information Scales C1 and C2, covering the period from the discovery of America to the Revolutionary War; Information Scales F1 and F2, covering the period from the Revolutionary War to the Civil War; and Information Scales K1 and K2, covering the period from the Civil War to the present time. These scales supplant the *Van Wagenen History Scales—Information and Character Judging*. The revised scales are suitable for grades five to twelve. The total time necessary is about forty minutes. The scoring is based on counting the errors and omissions of the three different groups of items into which each scale is divided.

195. Textbook Tests and Study Guides in American History. New and effective teaching and testing devices in the social studies have recently made their appearance for use in the secondary schools. These devices are of the work-book type. In these books unit problems are usually set up, not merely in the form of questions, but as challenging situations that invite pupil activity in their solution. They arouse the student's interest, stimulate him to investigation, and enable him to understand the social order in which he lives. They help him to recreate scenes and the people of other times. Furthermore, they afford him the satisfaction of completing definite pieces of work.

The organization of these work-books or study-guides is simple. Usually problems are outlined, and accompanying them are page references to all widely used textbooks, study

directions, questions, guides, new-type tests, suggestions for outside reading, space for the pupil's own written work, and space for illustrations, including current event clippings. Large accurate maps which are generally incorporated provide additional suggestions for pupil activity. The purpose of these devices is to create pupil interest and to increase power in the use of social science material. The authors of such devices realize that present-day teaching must emphasize the need for learning facts and formulating meanings and conclusions in situations in which the pupils themselves are active.

196. Measurement of Achievement in European and World History. The *American Council European History Test* is a comprehensive test covering a wide sampling of content suitable for use with high school and college students. Two equivalent forms are available. The working time for the test is ninety minutes.

The *Gregory-Owens Tests in Mediæval and Modern History* cover the major facts in mediæval and modern history from the fall of Rome to the present. The tests are specially designed for high school and normal school students but may be used in colleges giving a general survey course in this field. Two equal forms of the tests are provided. The working time for each form is forty-five minutes. The following summary shows something of the content of the test (Table 50).

TABLE 50

CONTENT OF GREGORY-OWENS MEDIÆVAL AND MODERN HISTORY TEST

Test Part	Field Measured	Number of Items
I	Miscellaneous Facts and Dates	10 questions
II	Association of Events	20 "
III	Historical Vocabulary	10 "
IV	Sequence of Events	12 "
V	Cause and Effect Relationships	12 "
VI	History of Middle Ages	10 "
VII	Modern History	30 "

The *Vannest Diagnostic Test in Modern European History* is a one hundred item test measuring five different aspects of history: (1) Time Sense, (2) Place Sense, (3) Evaluation of Facts, (4) Thought, (5) Information. The test exists in only one form and is administered without time limit.

197. **Tests for Civics and Government.** Measurement in the field of government and civics is limited to a few standardized tests representing two phases of the subject; namely, (1) information, and (2) attitudes.

Civic Information. The *American Council Civics and Government Test* is a comprehensive test for measuring achievement in high schools and colleges and for determining college entrance and placement. It has been scientifically constructed under the immediate supervision of the subcommittee on achievement tests of the *Central Committee on Personnel Methods* of the *American Council on Education.* In content the test is not limited to any one course, but is representative of the most widely used textbooks, and covers both fact material and reasoning. There are four parts to the test, employing four types of objective questions. Because of its comprehensiveness, it furnishes a very reliable and valid measure. There are two equivalent forms, A and B. The working time is 90 minutes. The correlation of the two forms is .88.

The *Almack Test in American Civics and Government* is issued in two forms designed and standardized for use in junior and senior high schools and normal schools. Each form consists of three parts comprising a total of 76 items.

The *Bear Test on the United States Constitution* is standardized for use in grades eight to twelve and in college. The test is designed to measure the student's knowledge concerning specific provisions of the constitution, and is validated by having each exercise refer directly to some provision in the constitution. The test does not measure knowledge of the history of the constitution nor its judicial interpretations. The reliability of the final edition was found to be .96. One class period is ample working time for the test.

The content of the *Brown-Woody Civics Test* has been selected as the result of carefully analyzing nine widely used

civics texts. No material is included that is not found in at least five of the texts. The test is made up of three parts: (1) civic vocabulary, (2) civic information, and (3) civic thinking. The manual of directions is excellent. End-of-year norms in percentiles are provided for the junior high school. Two forms of the test are available.

The *Burton Civics Test* is primarily an inventory test for use in grades five to nine. It is designed to reveal the civic information of pupils, their attitudes on civic questions, and their conceptions of proper modes of civic action when faced by life situations. Two forms are available. One of the unusual features of the test is that it provides norms for different economic groups. The author found that different types of norms were needed since children of different economic levels made significantly different scores.

The *Magruder-Chambers-Clinton American Civics and Government Test for High Schools and Colleges* has been constructed with the view to giving teachers in the field a valid and reliable measuring instrument in the fields of national, state, and local government. The arrangement of the test into Parts I and II and into smaller divisions permits the teacher to undertake a limited diagnosis of pupil difficulties. The reliability coefficient of this test is .85 based on 240 high school seniors. The working time is from 35 to 40 minutes.

Patterson's Test or Study Exercises on the Constitution of the United States are designed for use by any person interested in studying the Constitution, whatever his age or place in society. It is especially adapted for use in public and private schools, colleges, universities, clubs, and civic organizations. The test attempts to stress the constitutional provisions of most importance, as well as those items of civic information which should be the common knowledge of American citizens. It gives the instructor a modern teaching device and a convenient, accurate, and objective standard for measuring results. It presents to the pupil exercises carefully planned to arouse interest and produce a thorough knowledge of the Constitution. As a classroom study exercise, it may precede or accompany the regular instruction. The working time is approximately forty minutes.

The *Hill Civics Test* attempts to measure civic information and attitudes. Information is tested by the multiple response arrangement of four possible responses. Civic attitudes are also measured in the same manner. Many of the items of the information tests could well be called tests of vocabulary. Such terms as union, injunction, citizen, closed shop, corruption, city ordinances, etc., are included. In the attitudes section the pupil is asked to check the best of four possible answers with respect to the use of public property, driving a car without a license, obeying laws, value of an education, etc.

Hill-Wilson Test in Civic Action. This test deals with civic activities from the point of view of how, or what, to do. It does not measure civic action from the standpoint of behavior but rather for the purpose of discovering whether or not the pupil knows what to do in certain situations. The test is somewhat too limited in scope to secure a reliable measurement, but it is suggestive of a type which should prove useful in the field of civics and government.

198. Geography Tests. Many tests in geography have been prepared in the past few years but most of these are of the informational type. Few of the tests take into account the problem-solving aspects of social science instruction. Only one of the more recent tests makes any attempt to measure the underlying acquisitive abilities which the student must use in his mastery of geography.

The *Wiedefeld-Walther Geography Test* is a comprehensive test of the abilities, skills, and factual items which are represented in the modern aims of instruction in geography in grades four to eight. The content is based upon a careful check of forty-three courses of study and seven recent textbooks. The questions are grouped so that an analysis of the pupil's achievement may be made. Norms are based on nine thousand cases. The authors report a reliability of .82 in the fourth grade and .93 in the ninth grade.

The analysis in Table 51 gives a rather definite notion of the content of the test.

Other Geography Tests. Among the many other more recent geography tests and scales from which the teacher may find useful materials are the following, which are merely

TABLE 51

WIEDEFELD-WALTHER GEOGRAPHY TEST

Test Number	Subject	Number of Items	Testing Technique
I	Reading	16	Recognition
II	Organization	19	Best Answer
III	Map and Graph Reading	44	Recall-Identification
IV	Geography Vocabulary	30	Multiple-Choice; 4-Response
V	Geographical Relationships	30	Matching
VI	Place Geography	54	Identification Map Location

listed without comment: the *Gregory-Spencer Geography Tests*, the *Unit Scales of Attainment in Geography*, and the *New Stanford Geography Tests*.

199. **Miscellaneous Social Science Tests.** In this section brief descriptions of a few recent tests in miscellaneous fields of the social sciences are given.

The *American Council Economics Test* is a comprehensive test for high schools and colleges, similar in derivation, scope, and purpose to the American Council Civics and Government Test described above. There are two equivalent forms, A and B. The working time is 90 minutes. The correlation of the two forms is .82.

The *Burton Social Science-Civics Tests* are designed to indicate a pupil's ability to reason in social-civic situations. The test is issued in two forms, A and B, each form consisting of two parts. Part I covers social-civic information, and Part II deals with judgments in social situations. The tests show the relation of mere information to the ability to use that information in a social-civic situation. These tests are objective and cover the social-civic problems now being taught in these grades. They have been carefully standardized for grades 7 to 12, inclusive, and for normal schools.

The *Iowa Academic Tests in World History* are a part of

the series of unstandardized objective tests issued annually by the University of Iowa for use in the *Iowa Every-Pupil Testing Program.* In the 1934 series thirteen subject tests were prepared of which four were specifically in the field of the social sciences. These tests covered (1) American Government, (2) Economics, (3) United States History, and (4) World History. Each series of these tests is planned to parallel closely the instructional emphasis in this subject throughout the state. Thus for the state the tests comprising this series are particularly valid measuring instruments. The structure and content of the tests are changed from year to year so that it is impossible to give an exact description of the series.

The *Kepner Background Test in Social Sciences* covers material in history, geography, civics, etc., and aims to discover what background in these subjects the pupils bring into a course in social science. The test is made up of seven parts: (1) association of men and events, (2) literary background, (3) geographic concepts, (4) historical words and terms commonly used, (5) civic, social, and economic terms, (6) and (7) dates and chronology. It is available in two forms each containing 100 items using a variety of types of questions. Standards are provided for grades nine to twelve inclusive. No time limit is set, but usually not more than 40 minutes are needed.

The *Witham Hall of Fame Test* contains 63 names of great Americans in the Hall of Fame at New York University. A code of 15 classifications is given, e.g., author, educator, preacher, philanthropist, etc. The pupil is to place in front of each name the code letter to indicate under which classification each of the people belongs. Only one form of the test is available and it is designed for use in grades 7 to 12. No time limits are set for the test.

200. **Diagnosis and Remedy in the Social Studies.** Diagnosis in the social studies is difficult, first, because the knowledges, skills, and informations which pupils should learn are not too clearly identified; second, if the facts to be learned were known accurately it would still be impossible to determine whether the pupil functioned in his social relationships in a desirable manner merely because of his pos-

session of the informational elements revealed by a test. Diagnosis and remedy are often needed in those skills which are basic to successful work in the social studies. Achievement in these subjects requires much reading of the work-study type. Therefore, pupils, in order to achieve at acceptable levels, must possess many of the following work-study reading skills:

1. Knowledge of technical vocabularies employed in the social studies.
2. Reading comprehension with respect to adequate interpretation of social science content.
3. Ability to locate material readily — use of the index, library files, table of contents, maps, charts, etc.
4. Ability to outline.
5. Ability to summarize.

These skills are discussed in Chapter XIV along with other ways and means for corrective work in these important acquisitive skills.

Discussion Exercises

1. Try to define the field of the social studies in such a way as to clarify the objectives adequately for testing purposes.
2. Discuss the pros and cons of a unified social-science curriculum for the high school as contrasted with the traditional organization of the content by subjects.
3. To what extent do you believe the social science teacher is responsible for emphasizing the acquisitive skills lying back of achievement in the subject?
4. What procedure in the testing of problem-solving in arithmetic might be applied in the social sciences?
5. State four of the general outcomes of instruction in the social sciences.
6. Make a list of specific outcomes of instruction in history.
7. What are the three main types of social science tests as presented in this chapter?
8. In your judgment what is the relation of *fact* to *problem* solving in the social sciences?
9. What are the chief weaknesses in the problem-solving and the attitudes tests?
10. From the available tests in history, civics, and geography make a selection of one test in each subject suitable for use in the ninth grade. Defend your choices.

SELECTED REFERENCES[9]

Beard, Charles A., *A Charter for the Social Sciences.* Part I of Report of the Commission on the Social Studies, American Historical Association. New York: Charles Scribners' Sons, 1932.

Barton, William Alexander, *Outlining as a Study Procedure.* Columbia University Contributions to Education, No. 411. New York: Teachers College, Columbia University, 1930.

Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching,* Chapter XII. Boston: Houghton Mifflin Company, 1931.

Gilliland, A. R., Jordan, R. H., and Freeman, Frank S., *Educational Measurement and the Class Room Teacher.* New York: The Century Company, 1931.

Horn, Ernest, "Another Chapter on Tests for the Volume of 'Conclusions and Recommendations,'" *The Social Studies,* January, 1935.

Johnson, Henry, *An Introduction to the History of the Social Sciences,* Part II of Report of the Commission on the Social Studies, American Historical Association. New York: Charles Scribners' Sons, 1932.

Lindquist, E. F., and Anderson, H. R., "Achievement Tests in the Social Studies," *Educational Record,* April, 1933. Washington, D. C.: American Council on Education.

Mathews, C. O., *Grade Placement of Curriculum Materials in the Social Studies.* Columbia University Contributions to Education, No. 241, Teachers College, Columbia University, New York, 1926.

McCallister, James M., "Reading Difficulties in Studying Content Subjects." *The Elementary School Journal,* Vol. 31, November, 1930.

Phillips, Claude A., *Modern Methods and the Elementary Curriculum.* New York: The Century Company, 1931.

Pressey, L. C., "The Special Vocabularies of Public School Subjects." *Educational Research Bulletin of the Ohio State University.* Vol. 3, April, 1924.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

Stone, C. R., *Supervision of the Elementary School.* Boston: Houghton Mifflin Company, 1929.

[9] Teachers of the social studies are urged to read the various reports of the *Commission on the Social Studies* as they appear. Twelve reports are planned to appear from the press of Charles Scribners' Sons, New York.

Tormey, T. J., "The Effect of Drill Upon the Specific and General Comprehension of Historical Content." *Abstract of Doctoral Theses in Education* I, pages 153-182, Vol. IX, No. 1, University of Iowa Studies in Education, Iowa City, January, 1934.

Wilson, G. M., and Hoke, K. J., *How to Measure, Revised.* New York : The Macmillan Company, 1928.

Yoakam, G. A., *Reading and Study.* New York : The Macmillan Company, 1928.

CHAPTER XVI

MEASUREMENT IN MATHEMATICS

This chapter summarizes the following points involved in the measurement of junior and senior high school mathematics:

a. Significance of mathematics.
b. Junior and senior high school mathematics.
c. Arithmetic fundamentals.
d. Prognosis and achievement in algebra.
e. Measurement of abilities in geometry.
f. Advanced mathematics.

I. General Significance of Mathematics

201. **Significance of Mathematics.** Arithmetic and succeeding courses in mathematics constitute the intellectual background of exact thinking in the sciences and in the solution of many of the problems of modern society. For this reason arithmetic has long been recognized as a most important tool subject. Many think that failure to develop numerical concepts and to acquaint high school students with the rudiments of higher mathematics is to neglect a most important educational responsibility. We are born into a physical world in which quantity, shape, size, dimension, transformation and relationship play important roles. The universe is unquestionably mathematical. Whatever may be our background or our inclinations toward mathematics, our civilization, our sciences, our modes of living and of thinking have a mathematical core which we cannot escape. Mathematics, therefore, has a justifiable place in the secondary school curriculum. If properly taught and enriched it has the possibility of providing a type of training which is essential to every individual. It is considered by some to

be one of the permanent bases of both a practical and a liberal education.[1]

It is doubtless true, however, that in spite of its cultural values not all persons should be required to study mathematics. Many high school students find it unusually difficult and as a result an antagonism against the subject has arisen. From the standpoint of utility there are relatively few people aside from the specialists who need mathematics beyond the mere rudiments of arithmetic. There is also some question that the alleged cultural or disciplinary values of mathematics are not unproved assumptions. In any event, there is an apparently increasing tendency to make all mathematical work elective beyond the fundamentals as taught in the elementary grades.

202. **Mathematics in the Junior-Senior High School.** The elementary school, up to and including the sixth year, stresses the fundamental process of arithmetic computations. But life in our complex world sometimes requires the use of mathematical skills for which the work of the first six grades cannot fully prepare the individual. Junior-high-school arithmetic is planned to provide additional experience in the fundamental arithmetical processes through their use in solving problems arising out of the immediate and ultimate social experiences of this group. Other types of mathematical experiences in the junior-high-school curriculum aim to provide a gradual transition from elementary school arithmetic to senior high school mathematics and related subjects. The mathematical requirements which the senior high school must meet are of three kinds, namely:

1. Those arising through the demands of high school students who have no intention of going to college but who do require a background in practical and applied mathematics somewhat beyond the fundamentals of arithmetic.

[1] For excellent discussions of the educational values of mathematics see the following references:

Judd, Charles H., *Psychology of Social Institutions,* Chapter V, "The Psychology of Number." New York: The Macmillan Company, 1927.

Keyser, C. J., *The Human Worth of Rigorous Thinking.* New York: Columbia University Press, 1916.

Smith, David Eugene, "Mathematics in the Training for Citizenship," *Teachers College Record,* May, 1917.

2. Those arising through the requirements of special fields of instruction such as, manual arts, commercial arithmetic, physics, or chemistry.
3. Those arising because of college entrance requirements necessitating work in algebra, geometry, and advanced mathematics, i.e., solid geometry and trigonometry.

The foregoing statement of the major demands made on instruction in high school mathematics makes it clear that the problems of measurement in this field center around arithmetic, algebra, plane geometry, and perhaps a bit of solid geometry and trigonometry. In each of these subjects very specific and typical developments in measurement have taken place. While courses in general mathematics are becoming increasingly popular in junior high schools, relatively little has been done thus far toward the development of standardized tests in the subject. Accordingly, little or no space will be given to these tests in this chapter. In view of the very complete treatment of the measurement of arithmetic in the companion volume on elementary school tests only a brief consideration of the subject is presented. Algebra and geometry receive the major instructional emphasis in the high school and will accordingly be given the most attention here.

II. Arithmetic in the Junior-Senior High School

203. **Fundamentals of Arithmetic.** Despite the tendency toward the organization of general mathematics courses for junior high school, arithmetic is still taught as a separate subject in the seventh and eighth grades. Frequently, it is taught in later grades as a survey course for those who do not plan to enter college. Much of the instruction in these courses is reteaching. Inventory tests have shown repeatedly that in general pupils in the seventh and eighth grades do not have an adequate knowledge of the fundamental arithmetical principles and skills. Obviously, instruction in these review or survey courses should be based upon a careful diagnosis of each student's difficulties in the fundamental arithmetical operations which are most likely to be of use to him in solving the problems of everyday life.

One of the most recent and complete catalogues of the basic and teachable skills in the field of arithmetic is that reported in the *Twenty-ninth Yearbook of the National Society for the Study of Education.*[2] The student or teacher with special interests in arithmetic will of course wish to refer to this comprehensive report. Another very complete catalogue of arithmetical skills selected for teaching, testing, and remedial purposes is the one on which the *Compass Diagnostic Tests in Arithmetic*[3] are based. A summarization of this analysis is presented here to illustrate the extent to which such an analysis may be carried as well as to furnish a broad basis upon which to build diagnostic and remedial material in this field.

Arithmetic Skills

I. Fundamental Processes with Whole Numbers
   1. Basic Addition Facts
   2. Basic Subtraction Facts
   3. Basic Multiplication Facts
   4. Basic Short Division Facts
   5. Basic Vocabulary and Definitions of Arithmetic
   6. Basic Rules of Arithmetic
   7. Higher Decade Addition
   8. Column Addition
   9. Carrying in Column Addition
   10. Harder Subtraction
   11. Checking Errors in Subtraction
   12. Borrowing or Carrying in Subtraction
   13. Addition Used in Harder Multiplication
   14. Carrying in Addition Used in Harder Multiplication
   15. Complete Process of Multiplication
   16. Short Division Involving Carrying
   17. Multiplication, Addition, and Subtraction Used in Long Division
   18. Complete Process of Long Division

II. Fundamental Processes with Fractions and Whole Numbers
   1. Changing Fractions to Equivalent Forms
   2. Finding Common Denominators
   3. Reducing Fractions
   4. Addition of Fractions and Mixed Numbers

[2] *Twenty-ninth Yearbook of the National Society for the Study of Education.* Report of the Committee on Arithmetic, F. B. Knight, Chairman. Bloomington, Illinois: Public School Publishing Company, 1930.

[3] Ruch, G. M., Knight, F. B., Greene, H. A., Studebaker, J. W., *The Compass Diagnostic Tests in Arithmetic.* Scott, Foresman and Company, 1925.

5. Expressing Mixed Numbers as Improper Fractions
6. Fundamentals of Subtraction of Fractions
7. Reduction of Mixed Numbers
8. Cancellation in the Multiplication of Fractions
9. Reduction of Fractions and Mixed Numbers to Best Form in Answer
10. Multiplication of Fractions
11. Cancellation in Division of Fractions
12. Changing from Multiplication to Division Form
13. Fundamentals of Division of Fractions

III. Fundamental Processes with Decimals
1. Notation of Decimals
2. Changing Fractions and Mixed Numbers to Decimal Form
3. Changing Decimals to Fractions and Mixed Numbers
4. Fundamentals of Addition of Decimals
5. Fundamentals of Subtraction of Decimals
6. Place Values in Decimals
7. Pointing off in Multiplication of Decimals
8. Dividing Decimals by Pointing Off
9. Location of Decimal Points in Division
10. Changing Remainders to Decimal Form
11. Fundamentals of Division of Decimals

IV. Fundamental Processes with Denominate Numbers
1. Knowledge of Tables of Measure
2. Reducing in Denominate Numbers
3. Borrowing in Denominate Numbers
4. Addition of Denominate Numbers
5. Subtraction of Denominate Numbers
6. Multiplication of Denominate Numbers
7. Division of Denominate Numbers

V. Mensuration
1. Vocabulary of Mensuration
2. Mensuration of Plain Surfaces
3. Mensuration of Solids
4. Finding Areas and Volumes
5. Formulas Used in Mensuration

VI. Percentage
1. Fractional and Per Cent Relations
2. Decimal and Per Cent Relations
3. Expressing Areas in Per Cents
4. Fundamentals of Work in Percentage

VII. Interest
1. Vocabulary of Interest
2. Business Forms
3. Budgets

4. Computation of Interest
5. Computation of Discount
6. Use of Interest Tables

VIII. Problem Solving
1. Mastery of Fundamental Processes
2. Comprehension of Material Read in Problem
3. Knowledge of What Is Given in the Problem
4. Knowledge of What Is Called for in the Problem
5. Probable Answer to the Problem
6. Knowledge of Proper Processes to Use in Solving
7. Knowledge of Proper Order of these Processes
8. Recognition of the Correct Solution

204. **Survey and Diagnosis in Arithmetic.** Measurement in arithmetic is concerned chiefly with evaluating the factors of speed and accuracy in the fundamental processes and skill in solving verbal problems. Since it is not the purpose of this book to serve as a descriptive catalogue of tests, only a few of the typical arithmetic tests will be discussed.

Survey Tests in the Fundamentals of Arithmetic. Among the earlier tests for survey purposes the *Courtis Series B Arithmetic Tests* and the *Woody-McCall Mixed Fundamentals Tests* have rendered a long and useful service. The latter tests are still widely used in many quarters. They have the advantage of providing a quick cross-section of many arithmetical abilities, but because of the limited sampling of exercises at any given difficulty level the reliability of measurement is not particularly high. The computation section (Test 10) of the *New Stanford Arithmetic Test* is similar in many respects to the Woody-McCall fundamentals test, but is much longer, containing sixty carefully evaluated examples. The *Unit Scales of Attainment in Arithmetic* measure computations in the fundamental operations by means of a series of twenty-five scaled examples. Among the shorter tests, the *New Stanford Arithmetic Test* and the *Unit Scales of Attainment in Arithmetic* probably represent the most useful instruments.

The *Compass Survey Tests in Arithmetic* were designed to provide a reliable starting point for diagnostic measurement in connection with a constructive supervisory program in arithmetic. They furnish the necessary over-view of

arithmetic achievement, and point the way to needed diagnosis of a more detailed type. The tests are available in two equal forms, and in an elementary and an advanced examination. The Advanced Examination, which is standardized for use in grades four to eight inclusive, is in six parts covering the addition, subtraction, multiplication and division of whole numbers, fractions and decimals, as well as covering percentage and general problem solving. Sixty examples and exercises are presented in a thirty-five minute working period. Age and grade norms for the tests are based on returns from approximately 35,000 cases. The reliability of measurement for the tests is high.

Diagnostic Tests in Arithmetic. Among the diagnostic tests suitable for use in the seventh or eighth grades are two useful series each representing a rather specific point of view in diagnosis.

The *Brueckner Diagnostic Tests,* which cover whole numbers, fractions, decimals, and percentage, are really inventory exercises which make it possible for a sufficiently critical and analytical teacher to check through the pupil's work and discover the apparent causes of difficulty. The diagnosis thus becomes a matter of working out an individual analytical record for each child.

The *Compass Diagnostic Tests* represent another type of approach to diagnosis. This series consists of twenty tests covering the fundamental processes with whole numbers, fractions, decimals, percentage, arithmetical definitions and concepts, business forms, mensuration, and problem analysis and problem solving. The tests are essentially analytical in structure, the total process in each case being torn down one step at a time as a basis for the identification of the causes of weakness. These tests are designed for group measurement and diagnosis. They have been criticized because of their failure to take into account the speed of work of the pupils. It will be noted, however, that the several parts of the tests are given under reasonable time limitations and that as a result, test scores in terms of the number of examples right do represent speed at one hundred per cent accuracy. It is probable that no diagnostic test in any field is capable of indicating precisely *why* a skill breaks down,

but there is no more certainty that the teacher's interpretation of why the pupils encountered difficulty will be much more exact. The list of skills enumerated in the outline on pages 392 to 394 gives a very definite idea of the wide range of ability covered by these tests.

Commercial Arithmetic Tests.[4] Commercial arithmetic is taught for two purposes in many high schools: first, to provide a subject suitable to the commercial curriculum, and second, to provide an elective in mathematics which may be taken by students who either have met or are not at all interested in the college entrance requirements in mathematics. Among the recent developments in this field are the *Kinney Scales of Problems in Commercial Arithmetic.* The summary in Table 52 gives a definite idea of the structure and content of the test. In an effort to correlate commercial arithmetic with business activities, the makers of the test collected problems from merchants, business men and teachers. Every effort was made to select problems of real importance

TABLE 52

KINNEY SCALE OF PROBLEMS IN COMMERCIAL ARITHMETIC

Test	Part	Sub-Tests	No. Examples	When Used	Testing Time (min.)
A	1	Addition.......	40	After 10 wks.	5
		Subtraction.....	60		5
		Multiplication ..	60		5
		Division.......	40		5
	2	Aliquot parts...	12 of 5 problems	After 10 wks.	10
B		Written problems	10	End 1st Sem.	*
C		Written problems	10	End 2nd. Sem.	*

* No time limits set

[4] All tests mentioned in this volume are listed with publishers in the *Appendix*, pages 585 to 595.

in those business activities for which commercial arithmetic courses are intended to prepare students. The reliabilities of these tests are not particularly high. The tests are equipped with norms in terms of examples right and also in terms of a corrected score which according to the author gives a better discrimination than the raw score. T-scores may also be used if desired.

The *Newland Chart for Diagnosis of Illegibilities in Written Arabic Numerals* analyses those features of Arabic numerals which make for difficult or incorrect reading. This chart has been organized to facilitate remedial work aiming at the elimination of such illegibilities. It should be of decided value in public schools, business schools, and industrial firms that are attempting to make the writing of Arabic numerals more legible. It is based on the assumption that the first step in improving legibility should be the diagnosis of the specific difficulties followed by concentrated attention on the specific difficulties, rather than a blind effort at general improvement.

III. Measurement in Algebra

205. Objectives of Instruction in Algebra. Much attention has been given in the past few years to the restatement of objectives and in the reorganization of the subject matter of the secondary school. Mathematics, because of its cumulative character, is particularly sensitive to the reorganization of its content. Gradually a simplified program of instruction in algebra is being put into operation. Some of this simplification is due to the necessity of fitting the instruction to the maturity levels of the students. More of it is due, however, to the growing conviction on the part of teachers that the abstract symbolism of algebra has educational significance only if it is clearly understood and is applied in purposeful situations. Algebra, because of these factors, is now being organized around a relatively small number of concepts and practical topics. It includes the formula, the equation, and the graph, and as much purposeful technique as is necessary to make these topics effective for purposes of functional thinking and of problem solving. The special

purposes of the teaching of algebra may be summarized as follows:

1. To establish more adequately and to extend the meaning and significance of arithmetic.
2. To strengthen the pupil's power in computation, by much practice as well as by the development of devices useful in computation.
3. To develop the ability to interpret and to construct graphs correctly.
4. To develop understanding of the "language" of algebra.
5. To develop the ability to understand, solve, transform, and apply formulae.
6. To develop the equation and to apply it in the solution of problems of a wide range of interest, including large classes of problems often treated in arithmetic, as well as problems relative to geometry, to physics, and other natural sciences.
7. To furnish such material within its domain as may be needed in the later study of mathematics and of the various physical sciences.

206. **Measurable Qualities in Algebra.** An examination of the general objectives of instruction in algebra makes it clear that while there are numerous important specific skills, the general qualities of greatest significance deal with the development of speed and accuracy in the fundamental processes and in problem solving. While reasonable speed in performing the fundamental processes is desirable, the habit of mathematical accuracy must be developed in the student, if algebra is to fulfill its primary obligation. Speed of work, though important, ranks as secondary to accuracy and reasoning power. The majority of tests in algebra emphasize the mechanics of algebra in which speed and accuracy are primary elements. Much remains to be accomplished in the effective analysis and measurement of the problem-solving abilities.

207. **Tests in First Year Algebra.** Algebra tests differ considerably as to content and organization depending upon the purpose they are designed to serve. The selection of a test for use in the classroom depends upon the type of measurement desired. Representatives of the three general types of algebra tests, achievement, instructional, and prognostic, are briefly discussed in this section.

Achievement Tests. The *Breslich Algebra Survey Tests* provide an intensive measure of first semester and second

semester achievement of students' actual working knowledge of algebra and of their abilities to make a practical application of algebraic principles at the end of the first semester and the second semester of the course. Although designed primarily for survey purposes, these tests are divided into numerous parts for diagnostic or analytical purposes. The content of the tests (See Table 53) is based upon an analysis of the essentials common to eight leading textbooks in first year algebra.

TABLE 53

CONTENT OF BRESLICH ALGEBRA SURVEY TEST

Part	First Semester Test	Second Semester Test
I	Algebraic Concepts	Algebraic Concepts
II	Changing Algebraic Expressions to Simpler Forms	Algebraic Processes
III	Solving Equations	Solving Equations
IV	Deriving Equations from Problems	Problems
V	Formulas and Graphs	Fractions
VI	Factoring	Functional Relationships

The *Columbia Research Bureau Algebra Test* is available in two forms, A and B, and two separate tests. Test 1 is designed for use at the end of the first semester and Test 2 at any subsequent time. Each test is in two parts, Part I on mechanical operations, and Part II on solution of problems. Only essential operations are included, and the unimportant or seldom used ones are omitted. The two tests provide highly reliable and valid measures of students' ability and achievement in algebra. Each part of Test 1 can be given in forty minutes; Part I of Test 2, in thirty minutes and Part II in seventy minutes.

The *Douglas Standard Diagnostic Tests for Elementary Algebra*, as the name implies, are designed to provide diagnostic measures. It is doubtful, however, if these or any other tests in algebra are diagnostic in the strict sense of the term. The tests are arranged in two separate series. Series A is composed of four tests measuring (1) collection

of terms, (2) multiplication, (3) division, and (4) solution of simple equations. Ten exercises of increasing difficulty are included in each test. Series B consists of seven tests, as follows:

Test 1. Fractions
Test 2. Factoring
Test 3. Formulas and fractional equations
Test 4. Simultaneous equations
Test 5. Graphical representations and interpretations
Test 6. Square roots, exponents, and radicals
Test 7. Quadratic equations

Series A should be given near the end of the first semester and Series B near the end of the first year of algebra. The two series require thirty-four, and one hundred two minutes of working time respectively. For individual diagnosis the author of the tests recommends the use of the entire series of seven tests.

The *Iowa Academic Contest Every-Pupil Tests in First-Year Algebra* are described in connection with other measures of general educational achievement in Chapter XXII.

The *Iowa Unit-Achievement Tests in First Year Algebra* are designed to cover the basic phases of the entire year's work in beginning algebra. Two equivalent forms, A and B, are available. They are standardized for periodic use throughout the school year following the completion of instruction on a unit of work. The six tests comprising the series may be used instead of other types of examinations in this field. Tests 1, 2 and 3 cover the units of work normally taught in the first semester. Tests 4, 5 and 6 cover the second semester's work. No one of the tests is designed to be a semester final or term final examination.

Tests of this type possess several significant advantages over the usual survey type of test. In the first place, the content of the course of study may be more closely paralleled by such unit-type tests. This naturally results in greater assurance of validity in the tests. In the second place, since it is not necessary to confine the testing to a single period, it is possible to make tests somewhat longer thus permitting a more extensive sampling of the significant

items or skills. This results in increasing the reliability of the tests. In the third place, the periodic use of these unit-tests accomplishes one thing which end-of-the-year testing cannot do; it affords the teacher an opportunity to discover the weaknesses of pupils as they progress through the course. To discover pupil weaknesses at the end of the course is too late to hope to remedy them.

These tests are achievement tests, yet they offer many opportunities for diagnosis. The subject matter of each test is broken up into from three to six divisions or parts depending upon the extent and complexity of the unit. The scores on each of these parts are specific evidence of pupil mastery. Within each part there are from fourteen to forty specific reactions which represent a fair basis for the diagnosis of weaknesses in any particular division of the subject matter.

The tests and their respective divisions are as follows:

			Number of Exercises
Test 1.	Part 1.	Algebraic Language and Meaning	26
	Part 2.	Simple Equations	16
	Part 3.	Addition	20
	Part 4.	Subtraction	20
	Part 5.	Parentheses	14
	Part 6.	Multiplication and Division of Directed Numbers	14
		Total	110
Test 2.	Part 1.	Advanced Linear Equations	25
	Part 2.	Checking Equations	15
	Part 3.	Multiplication, Monomial and Binomial	40
	Part 4.	Verbal Problems	20
		Total	100
Test 3.	Part 1.	Division, Monomial	20
	Part 2.	Division, Polynomial	15
	Part 3.	Factoring	40
		Total	75
Test 4.	Part 1.	Equations Solved by Factoring	20
	Part 2.	Reduction of Fractions to Lowest Terms	16
	Part 3.	Fractions; Integral or Mixed Expressions	14
	Part 4.	Multiplication and Division of Fractions	20
	Part 5.	Addition and Subtraction of Fractions	20
		Total	90

Test 5.	Part 1.	Fractional Equations	25
	Part 2.	Formulas	20
	Part 3.	Problems Leading to Fractional Equations	30
		Total	75
Test 6.	Part 1.	Graphs	35
	Part 2.	Simultaneous Equations	25
	Part 3.	Elementary Quadratics	35
		Total	95

Instructional Tests. Instructional tests are different from the typical standardized test in that they consist of an extensive series of exercises, each covering a different unit of work and designed to be given at frequent intervals throughout the course. They provide, therefore, for a continuous program of instructional units, inventory and achievement tests, and well-timed remedial and corrective units. They should constitute an economical means of increasing the effectiveness of learning in the subject.

The *Schorling-Clark-Lindell Instructional Tests in Algebra* are excellent representations of this type of test series. They are organized to form a booklet of 52 standardized objective exercises covering the several skills in first year algebra. The booklet provides a means for diagnosis and measurement of ability at frequent intervals, and supplies material for daily drill to enable students to obtain the desired mastery of the important skills. Goals or standards for three levels of ability are set up. A uniform time limit makes it practical for students to progress according to ability even though the usual class organization is maintained. The tests may be used with any textbook.

Prognostic Tests. Teachers and high school supervisors have long been impressed with the excessive difficulty which pupils encounter in first year algebra. These difficulties are shown in two ways, (1) in the high percentage of pupil failure in the subject, and (2) in the large amount of extra help demanded by pupils outside of the classroom if failure is to be avoided. Because of the large amount of wasted time and effort, there is increasing interest in tests designed to predict pupil success in specific fields. In general, two

somewhat different types of procedure have been tried as a basis for prognosis. These are (1) the *learning* technique in which the aptitude of the student in the subject is measured in terms of the speed and accuracy with which he is able to acquire skills and information in the new field and respond to objective tests over the newly learned material, and (2) the *inventory* technique in which he reveals his aptitude in terms of reactions to specific exercises sampling into underlying skills upon which success in the subject depends. Of the tests described here the *Orleans Algebra Prognosis Test* is of the former type, while the *Iowa Algebra Aptitude*, the *Iowa Placement Test*, *Mathematics Aptitude*, and the *Lee Test of Algebraic Ability* are of the latter type.

The *Iowa Algebra Aptitude Test* is the result of an experimental study of a number of general tests for use as guidance instruments in this subject. The results of the experiment demonstrated the limited predictive power of one of the better group tests of mental ability so far as prognosis in first year algebra is concerned. It established clearly the necessity for a special type of prognostic test if adequate educational guidance in this subject is to be given. The summary in Table 54 gives an indication of the content and general character of this test.

TABLE 54

IOWA ALGEBRA APTITUDE TEST

Test Part	Name	Adm. time in min.	Possible Score	Reliability (N=223)
1	Arithmetic..............	12	30	.77
2	Abstract Computation....	8	25	.85
3	Numerical Series.........	12	42	.91
4	Dependence and Variation	3	10	.74
	Total............	35	107	.92

The data shown in Table 55 indicate clearly the relation of results on this test to those of other prognostic tests in

algebra and to algebra grades as well as to objective test scores in algebra at the end of the course. When it is considered that the *Iowa Algebra Aptitude Test* is a thirty-five minute test and that the composite of the two prognostic tests compared with it require a total of 106 minutes this becomes a significant statement of the relative validity of this aptitude testing technique.

TABLE 55

CORRELATION OF ALGEBRA APTITUDE TESTS

	Algebra Grades	Columbia Res. Alg.	Orleans Prognosis	Lee Test	Composite Lee-Orleans
Iowa Algebra Aptitude..........	.66	.76	.81	.80	*
Orleans Algebra Prognosis Test.....	.63	.72	*	.72	.96
Lee Test of Algebraic Ability...........	.59	.73	.72	*	.88
Composite of Lee-Orleans.......	.67	.78	.96	.88	*

The *Lee Test of Algebraic Ability* is an aptitude test based on the inventory type of approach. The four parts of the test measure (1) arithmetic problems, (2) analogies, (3) number series, and (4) formulas. The working time for the test is twenty-five minutes. The reliability reported for the test is .93.

The *Orleans Algebra Prognosis Test* is an aptitude test of the learning type. It gives a basis for predicting the pupil's success in the subject by measuring the speed and accuracy with which he is able to learn novel material of the sort encountered in algebra. The test contains eleven simple lessons with a test on each covering fundamental principles and essential skills in learning algebra. An arithmetic and a summary test are also included. The working time for the

complete test is eighty-one minutes. The reliability is not given in the examiner's manual.

208. College Placement and Prognosis Tests. The *Iowa Placement Examinations, Mathematics Aptitude* and *Mathematics Training,* have had a wide usefulness in discovering the aptitudes of college students and in determining the effect of preliminary training in this field. The aptitude series is designed to measure those factors which contribute to subsequent success in the subject. The training series measures the effect of previous training in subjects which are considered prerequisite to the work at hand. The aptitude series assumes no special training in the subject. The training series assumes one year of high school mathematics or one semester of college training. The reliabilities of the examinations ranges from .87 to .94 for beginning college students. The *Mathematics Aptitude Test* and the *English Training Test* of this series have been used with much success to supplement other tests of general mental ability in freshmen testing programs at the college level.

IV. Plane Geometry

209. Objectives of Instruction in Plane Geometry. The more important general objectives of high school geometry may be summarized as follows:

1. The development of logical reasoning ability.
2. The development of an appreciation of the utility and beauty of geometrical forms.
3. The familiarization of the student with the properties, mensuration, and relationships of common geometric forms.
4. The development of an understanding and an appreciation of deductive proofs.
5. The creation of an understanding of spatial concepts and relations.
6. The establishment of habits of precision and accuracy.
7. The development of an appreciation of the part geometry has played in the history of civilization.[5]

210. Possibilities of Measurement. The general nature of the objectives of instruction in geometry makes it neces-

[5] From *Instruction in Mathematics,* Monograph No. 23 of the National Survey of Secondary Education, U. S. Office of Education, 1932.

sary to state these outcomes in very definite form before the measurable qualities in the subject can be identified. The accompanying summary of knowledges, skills, and abilities is at the same time a catalogue of those elements of plane geometry which lend themselves most readily to objective measurement. This list is a compilation based upon numerous statements of objectives, textbooks, tests and courses of study.

1. Ability to visualize or think in terms of visual imagery.
2. Ability to read geometrical material, theorems, problems, etc.
3. Ability to use efficiently the basic arithmetical processes required.
4. Ability to apply the algebraic processes required.
5. Ability to state what is given in the theorem or problem.
6. Ability to state what is to be proved in the theorem or problem.
7. Ability to state in proper and logical order the essential steps in the proof or solution of the problem.
8. Ability to recall or select additional facts necessary for the proof or solution.
9. Ability to detect errors in reasoning in a proof or solution.
10. Ability to detect errors in the mathematics of the proof or solution.
11. Ability to draw figures required by a proof or problem.
12. Ability to detect errors in figures required by problem.
13. Ability to do certain constructions.
14. Ability to detect errors in the steps in a construction or in the construction itself.
15. Ability to apply geometric principles to real life problems.

Not all of these elements of accomplishment in plane geometry are measured directly in any one series of tests although many may appear incidentally in connection with the total process of proving a theorem or solving a problem. While everyone recognizes the importance of measuring the pupil's power to reason logically from given data as a most important outcome in geometry it must nevertheless be admitted that the most effective measurement of such reasoning power is revealed through his success in the application of specific abilities and skills.

211. **Measures of Achievement in Plane Geometry.** The nine tests of achievement in plane geometry described in this section are representative of the more recent and more useful types of tests in this field.

The *Breslich Geometry Survey Tests* are available in a

first semester and a second semester test, each in two approximately equivalent forms. The items were carefully selected and validated, and the form in which they are stated was determined experimentally. The author indicates that the following outcomes are measured by the tests:

1. Ability to understand and use intelligently the fundamental concepts of the geometry course.
2. Knowledge of the fundamental constructions of geometry.
3. Familiarity with the most frequently used relationships and the ability to express them in the most convenient symbolic form.
4. Sufficient reasoning power to apply the postulates of geometry and draw inferences from given or known facts.
5. Ability to recognize and to use important geometric principles when they are found in practical situations.

Since the tests are largely self-administering, they also measure some aspects of the student's ability to secure meaning from the printed page and specifically to read directions relating to geometric material. The various parts of the test are administered separately thus adding considerably to the diagnostic value of the results. The working time for the first semester test in thirty minutes; for the second semester test, fifty-one minutes.

The *Columbia Research Bureau Plane Geometry Test* is available in two equivalent forms, A and B. The content of the test covers the more important topics of a one-year high school course and is made up of problems and of questions concerning many aspects of the subject. The test has proven to be a reliable and valid measure of the student's knowledge and ability in plane geometry at the end of the course. It may be used as a final examination in high schools, or as an entrance test in college. A supplement to the manual of directions makes provision for giving a more extended examination when that is desired. The working time is sixty minutes. The reliability coefficient is .93.

The *Lane-Greene Unit Achievement Tests in Plane Geometry* are standardized for periodic use throughout the school year following the completion of instruction on each unit of subject matter. The items were carefully validated to conform to the recommendations of current committees on geometry content. The vocabulary, theorem and prob-

lem content of the tests is in accordance with practices as found in eighteen textbooks. Data showing the general plan of organization of the tests, and certain other critical information are given in Table 56.

TABLE 56

LANE-GREENE UNIT ACHIEVEMENT TESTS IN PLANE GEOMETRY

Test	Subject	Possible Score	Administration Time (min.)	Reliability (N=200)
1	Fundamental Ideas........	34	35	.83
2	Parallel lines and triangles..	60	38	.89
3	Rectangular figures........	59	38	.90
4	The circle................	45	38	.83
5	Proportion and similar polygons....................	45	36	.86
6	Areas of polygons.........	45	36	.86

The advantages of unit-tests of this type are presented by the authors as follows:

1. *The tests themselves more accurately parallel the content of the course of study upon which instruction is based.* This results in greater assurance of validity in the tests.
2. *The individual tests can be somewhat longer,* thus permitting a more extensive sampling of the important skills or items of information, *which in turn increases the reliability of the resulting measures.*
3. *The practice of using unit-achievement tests on practical units of instructional material periodically throughout the year insures the teacher an opportunity to identify the specific difficulties of individuals early enough in the course to give corrective instruction on the weaknesses.* End-of-the-year testing obviously does not do this.

While these tests are measures of achievement and are not primarily diagnostic, they offer considerable opportunity for diagnosis. The content of each test except the first is divided into four parts representing important aspects of accomplishment in each of the units. The first three tests are on units of work uniformly taught in the first semester

of the course in plane geometry. The last three tests cover the work commonly given in the second semester.

The *McMindes Plane Geometry Tests* are designed primarily for the measurement of achievement at the completion of the course in plane geometry. There are two forms of the test, A and B, each consisting of three divisions, viz: true-false, multiple-choice, and problems. There are a total of 136 test elements included. The true-false and multiple-choice divisions are probably more thought-provoking than the elements ordinarily used in this type of test. The problems section is devised to test geometric principles and to minimize the mechanics of solution. The test is very easy to give and score. The working time for the test is sixty minutes.

The *Orleans Plane Geometry Achievement Test* is a comprehensive survey instrument. It is available in two forms, A and B, comprising two tests each. *Test* 1, for first semester, covers Books I and II except loci; *Test* 2, for second semester, covers Books III, IV, and V and loci. There are five parts to each test, each part dealing with a different phase of the subject, including geometrical reasoning, computation, completing proofs, giving diagrams and hypotheses and conclusions, and original solutions. The tests make possible accurate analysis of students' difficulties and causes of failure. This should be of real aid to the teacher in planning instructional work of a remedial character, and in classifying students. Each test can be given in a typical high school period. The reliability coefficients for the tests are .85 for Test 1 and .71 for Test 2.

The *Perry Geometry Tests* comprise five tests designed to cover the year's course in plane geometry. The general plan of organization of the content is quite similar to that followed in the Lane-Greene Unit Achievement Tests except that the units are considerably over-balanced for what is normally the second semester's work. The working time for the tests is thirty-five minutes each.

The *Renfrow Diagnostic Tests* in Plane Geometry are intended to provide a diagnostic measure but they are of limited value for that purpose until after the latter part of the first semester's work or the beginning of the second

semester's work has been completed. Used at that time, the tests reveal with fair accuracy class weaknesses in any phases of geometry sampled by the various tests. The content of the tests is indicated by their titles, as follows:

Test 1. Construction and Locus Problems
Test 2. Construction in Theorems
Test 3. Definitions
Test 4. Demonstration of Theorems
Test 5. Original Problems

The *Schorling-Sanford Achievement Test in Plane Geometry* first appeared in 1921, but since that time has been revised three times. Each of the two forms of the complete test is composed of five tests each containing twelve exercises. The test is intended as an end-of-year measure. The content is carefully validated in accordance with the best current practices in this subject. The five tests comprising each test booklet are as follows:

Test 1. Completion sentences (information concerning geometric facts)
Test 2. Conclusions from given data
Test 3. Judging conclusions
Test 4. Demonstration of theorems
Test 5. Original problems

The *Webb Plane Geometry Test* is available in only one form which is organized into a test booklet comprising five parts. In Part I a figure is given and the student is asked to complete statements regarding this figure. Parts II, III, and IV test the student's knowledge of geometric terms and theorems. Part V requires the student to construct geometric figures. The testing time is forty minutes. The reliability of the test is given by the author as .87.

212. **Instructional Tests in Plane Geometry.** The general plan of organization of instructional tests in plane geometry is similar to that used in all such tests. Only one series of such tests is described here as typical of these instruments. The *Bishop-Irwin Instructional Tests* comprise a seventy-two page booklet containing a series of thirty-six standardized tests each covering a unit of work in the regular high school

course and including all phases of the subject. By giving the tests at approximately weekly intervals the teacher can determine accurately the student's achievement in the work just completed, and thus have a reliable basis for remedial teaching and review to fix essentials and for adaptation of the instruction to individual needs.

213. **Prognosis in Geometry.** The modern high school enrolls thousands of students who are not capable of doing satisfactory work in such an abstract subject as plane geometry. Much wasted effort and time can be saved by both teachers and pupils through the proper use of standardized instruments capable of revealing the individual students who are not fitted to pursue the study of the subject. Two types of aptitude tests for this purpose have been developed thus far in the field of plane geometry. These two types are represented by the *Iowa Plane Geometry Aptitude Test* and the *Orleans Geometry Prognosis Test.* The general plans of construction of these two tests are the same as those followed in the prognosis of algebra abilities.

The *Iowa Plane Geometry Aptitude Test* is designed to furnish a reliable prognosis of pupil accomplishment in this field prior to any study of the subject. The test utilizes the inventory rather than the learning technique as the basis of the prognosis. The basic abilities measured, together with other significant data concerning the tests, are shown in Table 57.

TABLE 57

IOWA PLANE GEOMETRY APTITUDE TEST

Part	Test Name	Possible Score	Adm. Time (min.)	Reliability (N=414)
1	Reading Geometry Content	20	9	.84
2	Algebraic Computation.....	17	9	.84
3	Algebraic and Arithmetic Reasoning..............	18	13	.83
4	Visualization.............	20	13	.91
	Total.............	75	44	.90

As evidence of the reliability of the test the authors submit the fact that a correlation of .71 was found for results from 131 students on this test and objective test scores at the end of the first semester. A correlation of .59 was found on 146 cases for the test and the average of first and second semester grade marks. The correlation of the *Iowa Plane Geometry Aptitude Test* with the summation of three unit achievement tests (Lane-Greene, Tests 4, 5, and 6) representing second semester work was .72.

The *Orleans Geometry Prognosis Test* is similar in its general plan to the *Orleans Algebra Prognosis Test* described on page 404. In general, it utilizes the learning procedure as the basis for prognosis. It consists of several simple lessons in geometry with a test on each lesson, involving essential skills and fundamental principles. It is intended to be used before the study of the subject is begun. The working time for the test is seventy minutes.

V. Solid Geometry and Trigonometry

Solid geometry and trigonometry are generally considered to be advanced mathematics and are taught as separate and distinct from each other and from other mathematics courses. For this reason they are treated in a separate section.

214. **Solid Geometry.** The instructional purposes of solid geometry are given expression in the following recommendation from the Report of the National Committee on Mathematical Requirements:[6]

> "The aim of the work in solid geometry should be to exercise further spatial imagination of the student and to give him both a knowledge of the fundamental spatial relationships and the power to work with them. It is felt that the work in plane geometry gives enough training in logical demonstration to warrant a shifting of emphasis in the work on solid geometry away from this aspect of the subject and in the direction of developing greater facility in visualizing spatial relations and figures, representing such figures on paper and in solving problems in mensuration."

In general, students of solid geometry are expected to reveal through their solution of numerical originals, familiarity

[6] Report of the National Committee on Mathematical Requirements, *The Reorganization of Mathematics in Secondary Schools.*

with the fundamental formulas of plane geometry, with the use of logarithms, with the use of natural and logarithmic functions (including sine, cosine, and tangent). Furthermore, they are generally expected to reveal familiarity with the solution of triangles and with the formulas of solid geometry. The subject is generally elective, and the number of students electing it is not as large as for algebra and plane geometry. Perhaps this is one of the reasons why little attention has been given to the construction of standardized tests for this subject. Only two will be described briefly below.

The *Seattle Solid Geometry Tests* are made up of nine sub-tests (to be used semi-monthly during the semester) and of two scales entitled "Scale A, Final Examination," and "Scale B, Preliminary to the Final Examination." The nine sub-tests consist of a total of three hundred true-false items grouped equally around the following sub-heads:

1. Lines and Planes
2. Lines and Planes
3. Dihedral and Polyhedral Angles
4. Prisms and Parallelopipeds
5. Pyramids
6. Cylinders and Cones
7. Spheres, Cylinders, and Cones
8. Spherical Polygons and Spherical Areas
9. Spherical Volumes and Areas

Cumulative scores on the nine tests throughout the semester provide relative achievement and diagnostic data concerning each pupil. The carefully scaled *Scale A, Final Examination*, consists of sixty-three items, seven items from each of the nine units listed above but not duplicating any of the items previously used. *Scale B, Preliminary to the Final Examination* is an equivalent duplicate of *Scale A, Final Examination* and is provided for research and survey work. Since some of the items of *Scale B* are included in the nine sub-tests, it should not be used as a final examination where the sub-tests have previously been administered. A time limit of twenty to twenty-five minutes each for the nine tests is required and forty-five to fifty minutes for Scale A or Scale B.

The *American Council Solid Geometry Test* was devised for measuring achievement in high schools and colleges, and for college entrance and placement testing. In content the test is not limited to any one course, but is representative of the most widely used textbooks in the subject. There are two parts to the test, one covering fundamentals, and one covering problem solving. The test is available in two forms, A and B. The working time for the test is sixty minutes.

215. **Trigonometry.** Measurement in trigonometry is concerned with two major elements: (1) an understanding of the fundamental proposition of trigonometry, and the familiarity with the fundamental relations between the ratios; (2) the ability to solve problems involving the applications of such understandings. A brief description of one test in Trigonometry is presented here.

The *American Council Trigonometry Test* for high schools and colleges is available in two forms, A and B. It was devised for the Committee on Personnel Methods of the American Council on Education. This test is similar in purpose to the American Council Solid Geometry Test described on this page. In five parts, it covers the fundamentals including trigonometric functions, logarithms, and problem solving. The working time for the test is eighty minutes.

Discussion Exercises

1. In your judgment what accounts for the fact that algebra and geometry have not been subjected to the same detailed analysis and extensive measurement as is true of arithmetic?
2. Identify ten or more of the specific skills in upper grade arithmetic which appear to lend themselves to measurement and remedial treatment.
3. To what extent do you find that we are justified in depending upon transfer from algebra to aid in learning in geometry?
4. Why are there no adequate diagnostic tests in algebra or geometry?
5. Make a comparison of any two selected algebra or geometry tests by comparing complete lists of specific skills measured by each.
6. Show how the basic steps in problem solving in algebra and geometry closely parallel the steps in the thinking process.

7. Evaluate the two types of techniques which are proposed as the basis for prognosis in algebra and geometry.
8. What procedures can you suggest for the improvement of initial instruction on problem solving in algebra and geometry?
9. What are the chief advantages of the so-called unit-achievement type of tests?
10. Outline a testing and remedial program in some one field of high school mathematics indicating where possible your first and second choices of material. Give reasons for your choices.

Selected References

Breslich, Ernest R., *The Teaching of Mathematics in Secondary Schools.* Chicago: The University of Chicago Press, 1930.

Brueckner, L. J., *Diagnostic and Remedial Teaching in Arithmetic.* Philadelphia: John C. Winston Company, 1930.

Brueckner, L. J., and Melby, E. O., *Diagnostic and Remedial Teaching.* Boston: Houghton Mifflin Company, 1931.

Greene, Charles E., and Buswell, G. T., "Testing, Diagnosis, and Remedial Work in Arithmetic." Report of Society's Committee on Arithmetic. *Twenty-Ninth Yearbook, National Society for the Study of Education.* Bloomington, Illinois: Public School Publishing Company, 1930.

Greene, H. A., and Lane, Ruth O., "Validation of New Types of Plane Geometry Tests," *Mathematical Teacher,* Vol. 22: 293-301, May, 1929.

Greene, H. A., Studebaker, J. W., Knight, F. B., Ruch, G. M., *Economy Remedial Exercises in Problem Solving in Arithmetic.* Chicago: Scott, Foresman and Company, 1926.

Judd, Charles Hubbard, *Psychology of Secondary Education.* Chapters 6, 7, 8. Boston: Ginn and Company, 1927.

Knight, F. B., Report of Arithmetic Committee, as Chairman, *Twenty-Ninth Yearbook of the National Society for the Study of Education.* Bloomington, Illinois: Public School Publishing Company, 1930.

Knight, F. B., and Buswell, G. T., *National Council of Teachers of Mathematics, Second Yearbook.* New York: Bureau of Publications, Teachers College, Columbia University, 1927.

Langlie, T. A., "A Comparison of Aptitude and Training Tests for Prognosis," *Journal of Educational Psychology,* Vol. 19: 658-65, December, 1928.

Lide, Edwin S., *Instruction in Mathematics.* Monograph No. 23, National Survey of Secondary Education, Bulletin No. 17.

Washington, D. C.: United States Department of the Interior, Office of Education, 1932.

Ligda, Paul, *The Teaching of Elementary Algebra.* Boston: Houghton Mifflin Company, 1925.

McCormick, Clarence, *The Teaching of General Mathematics in the Secondary Schools of the United States.* Teachers College, Columbia University. Contributions to Education, No. 386, Ch. V. New York, 1929.

Morton, R. L., *Teaching Arithmetic in the Intermediate Grades.* New York: Silver, Burdett and Company, 1927.

Odell, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

Reeve, W. D., *A Diagnostic Study of the Teaching Problems in High School Mathematics.* Boston: Ginn and Company, 1926.

Sixth Yearbook of the National Council of Teachers of Mathematics.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

Smith, D. E., and Reeve, W. D., *The Teaching of Junior High School Mathematics.* Boston: Ginn and Company, 1927.

Symonds, P. M., *Measurement in Secondary Education.* New York: The Macmillan Company, 1927.

Uhl, Willis L., and Others. *The Supervision of Secondary Subjects.* Chapter 3. New York: D. Appleton and Company, 1929.

Upton, C. B., *Tests in Geometry.* In *The Reorganization of Mathematics in Secondary Education* published by the Mathematical Association of America, 1923. (Obtainable from J. W. Young, Dartmouth College, Hanover, New Hampshire.)

Young, J. W. A., *The Teaching of Mathematics in the Elementary and Secondary School.* New York: Longmans Green and Company, 1929.

CHAPTER XVII

MEASUREMENT IN THE SCIENCES

The purpose of this chapter is the discussion of the following points involved in the measurement of instruction in the sciences:

a. Significance of the sciences.
b. Scope of science teaching in the secondary schools.
c. Objectives of instruction in the sciences.
d. Measurable qualities in science.
e. Tests of achievement in the sciences.
f. Instructional tests in the sciences.
g. Analysis and remedy.

I. General Significance of the Sciences

216. **Science and Human Progress.** The influence of science on human progress has been widespread and real. Through its teachings superstitions and erroneous beliefs of long standing have been corrected. Science has shrunk the earth by speeding up transportation and communication. It has brought a new era of health and new controls over disease. It has made us masters of energy and machines which in a comparatively short period of time have practically transformed our manner of living. The present mechanical age is the outgrowth of this control, and in many cases has almost completely upset prevailing social and economic standards. In this respect, at least, it is apparent that not all of the contributions of science to humanity have been without their disturbing aspects.

217. **Instructional Outcomes in the Sciences.** A superior quality of science instruction should be expected as a matter of course in a scientific age such as the present. It is surprising, therefore, to find that progress in the clarification of objectives of instruction, the selection of the subject-matter content, the improvement in teaching and in testing methods

and materials in recent years has been relatively slight, in spite of the great practical value of the subject and its natural appeal to the interests of students of all ages. Significant educational values will be attained if the student, as a result of instruction in the sciences, shall acquire: (1) an ability to utilize the findings of science that apply to his own daily experiences; (2) an ability to interpret correctly the natural phenomena of his environment; and (3) an appreciation of scientific attitudes and methods used by creative workers in the fields of science to the extent of applying them in the solution of his own social, civic, and health problems.

To supplement these somewhat general objectives of science instruction, an attempt has been made to identify some of the larger and more important scientific concepts. The Committee on the Teaching of Science has suggested the following principles and generalizations which it considers of value for guidance in selecting and organizing content for science teaching:[1]

1. The sun is the chief source of energy for the earth.
2. Through interdependence of species and the struggle for existence, a balance seems to be maintained among the many forms of life.
3. The earth's position in relation to the sun and moon is a determining factor of life on earth.
4. All life comes from life and produces its own kind of living organism.
5. Matter and energy cannot be created or destroyed, but may be changed from one form to another.
6. Species have survived because of adaptations and adjustments which have fitted them to the conditions under which they live.
7. The energy of solar radiation is continually working changes in the surface of the earth.
8. There have been profound changes in the climate, not only of certain regions, but also of the earth as a whole.
9. The evolution of the earth has come as a result of natural forces.
10. Units of time are defined by the earth's movements in relation to the sun.
11. All life has evolved from simple forms.
12. The earth seems very old when measured in the ordinary units of time.
13. Distances in space seem extremely vast when compared with distances on earth.
14. The physical environment has great influence on the structural forms of life and on plant and animal habitats.

[1] *Thirty-first Yearbook, Part I.* National Society for the Study of Education, 1932.

15. Man can modify the nature of plant and animal forms through application of his knowledge of the laws of heredity.
16. There is a great variety in the size, structure, and habits of living things.
17. There are processes that go on within an organism that are vital to its continued existence.
18. Chemical and physical changes are manifestations of energy changes.
19. There are fewer than one hundred chemical elements.
20. Every substance is one of the following: (a) a chemical element, (b) a chemical compound, (c) a mechanical mixture.
21. Certain material substances and certain physical conditions are limiting factors to life.
22. Light is a limiting factor to life.
23. Sound is caused by waves which are produced by a vibrating body which can affect the auditory nerves of the ear.
24. Gravitation is the attractive force that influences or governs the movements of astronomical bodies.
25. Machines are devices for accomplishing useful transformations of energy.
26. Any machine, no matter how complicated, may be analyzed into a few simple machines.
27. The properties of the different elements depend on the number and arrangement of the electrons and protons contained in their atoms.
28. All matter is probably electrical in structure.
29. The applications of electricity and magnetism in the home and in industry have revolutionized the lives of many people.
30. Heredity determines the differences between parents and offspring as well as the resemblances.
31. The kinetic energy of the molecules determines the physical states of matter.
32. The gravitational attraction between the earth and a mass of unconfined gas or liquid causes the pressure of the liquid or gas on the surface of the earth.
33. Liquid or gas pressure is exerted equally in all directions.
34. Chemical changes are accompanied by energy changes.
35. A change in rate or direction of motion of an object requires the application of an external force.
36. Radiant energy travels in straight lines through a uniform medium.
37. Electricity is a form of energy that results from disturbing the position or the regular paths of electrons.
38. In a chemical change a quantitative relationship exists between the amounts of substances reacting and the amounts of substances that are the products of the reaction.

This list of principles and generalizations gives an idea of the breadth and depth of the field and also a clue as to the types of objectives to be sought in science instruction.

Apparently, the aim of science teaching as expressed in these generalizations is contributory to the general aim of education, the enrichment of the life of the individual. As a matter of fact, the objectives are two-fold: (1) the functional understanding of the major generalizations of science, and (2) the development of the desired associated scientific attitudes.

218. **Scope of Instruction in the Sciences.** The broad range of subject matter which comprises the school sciences complicates measurement in this field of instruction. Elementary school science now commonly includes nature study and physiology and hygiene, with some overlapping of these with physical geography, biology, and general science in the junior-high-school grades. In general the science work in the junior-high-school level should be based upon and comprehend the science of the first six grades. Furthermore, it should serve as an introduction to the special sciences of the high school for those pupils who continue in school beyond the ninth grade.

The content of junior-high-school science should provide those conditions which will enable the student of this age to satisfy his curiosity concerning phenomena and applications in the field of science, and to make contact with the more important practical aspects of science in everyday life. The senior high school sciences are apt to be quite specific and involve sampling much more deeply into limited fields. The sciences commonly emphasized in the senior high school are biology, chemistry, and physics.

II. Possibilities and Limitations of Measurement

The real evidences of accomplishment in the sciences are to be found in the development and direction of the pupil's interests, attitudes, appreciations, skills, habits, ideals, and actions in the fields of the sciences. The ideal way to determine the changes which are effected in the pupil as a result of studying a unit of science would be to measure the increment of desirable activities which he can and does produce as a result of this study. However, thus far no such tests or suggestions for making such tests are available. Almost

nothing has been done to measure the acquisition of laboratory techniques and abilities although this should not be unduly difficult. The average science course apparently attempts to accomplish little more than to give a knowledge of the names of a few of the common animals, plants, and physical objects, and an acquaintance with a few of the simpler natural phenomena, without any very definite purpose appearing in the accumulation of such information. Test construction in the sciences has to a large degree paralleled these purposes.

219. **Measurable Qualities.** Four major types of measurable qualities may be designated in the sciences. These are discussed in order in this section.

Facts. Most existing science tests tend to emphasize information and knowledge as the goal of study. It is too often assumed that knowledge is a positive index of satisfactory modes of adjustment. This assumption, of course, is only partially defensible. Mere knowledge gives no assurance of subsequent proper reaction. But, in so far as knowledge is essential to adjustment, its proper worth should not be discounted. Accordingly, measurement of the pupil's knowledge of scientific facts is to that extent valid and defensible.

Relationships. Facts in science are the vehicles for thought. The understanding of the general relationships of facts is the knowledge deemed most important. It is these generalized ideas which students should acquire. Tests should, therefore, so far as possible, measure the relational aspects of science.

Problem Solving. Problem tests in science call for the application of knowledge, and may demand one or more types of scientific thinking. Similarly, test items which involve the interpretation of new situations demand more than mere recall and, thus, are measures of ability to use scientific knowledge or judgment. Such test items should find a more extensive place in testing procedures.

Attitudes and Interests. Little progress has been made in the measurement of the interests and attitudes of pupils with respect to scientific materials and phenomena. The task is a difficult one because as yet the interests and the at-

titudes deemed desirable have not been defined very clearly. Part of the delay in preparing such tests is probably occasioned by the present emphasis on subject matter, and the inherent difficulty of applying precise measures to elements so intangible as interests and attitudes.

220. Measurement in the Elementary Sciences. Such elementary science tests as are available are largely memory tests of factual materials.

Nature Study. No standardized tests in nature study are known to the writers. It is quite probable that for the immediate future, at least, the tests in this field will be mainly those constructed by the instructors in each local school system. The construction of standardized tests in nature study must wait upon a more universal agreement as to the aims and purposes of teaching nature study, a statement of more representative criteria for the content of the course of study in this subject, and a more definite identification of the minimum essentials of the course.

The elementary science section of *The Modern School Achievement Tests* is practically the only standardized test which approaches this subject-matter field. The test itself contains fifty multiple-choice exercises requiring twelve minutes for administration.

Hygiene. The standardized tests in physiology and hygiene are limited in number and quality. One of the most usable of the tests in this subject is Test 8 of the *New Stanford Achievement Tests, Advanced.*

General Science. The most significant movement in the development of a unified course in science has taken place in grades seven, eight, and nine. In the senior high school grades this General Science course has not been well received, and the separate subject-matter courses of biology, botany, chemistry, and physics are still taught in most high schools. A number of highly useful standardized tests in general science are available. These are discussed and described in a later section of this chapter.

III. MEASUREMENT IN GENERAL SCIENCE[2]

221. Objectives in General Science. General science is primarily organized to emphasize the study of the environment. To this end the following specific aims were prepared by the New York Regents[3] to guide teachers of the subject in their instructional emphasis:

1. To increase the pupil's appreciation and awareness of the factors of his environment.
2. To develop the pupil's powers of observation through a carefully directed study of the common phenomena of his immediate environment.
3. To give the pupil training in the use of the scientific method of problem solving.
4. To stimulate the pupil so to organize his common experiences that he shall make a more direct and purposeful use of his activities.
5. To foster in the pupil a confidence and belief in the knowledge he gains by the use of the scientific method to the extent that it will influence his daily conduct through the formation of good habits.
6. To develop the pupil's practical knowledge of personal and community hygiene.
7. To give the pupil a working knowledge of some of the specific ways in which the facts, principles and laws of his environment are related to and used in the world's work in order that he may more fully appreciate the vital place of science in the community and its bearing upon national progress.

Of these objectives, the ones involving factual matter or knowledge in the sciences are the most readily measured. Relatively little success has accompanied attempts to measure objectively the student's ability to utilize scientific methods, or even to assume a scientific attitude of mind. It is doubtful if the matter of interests implied in bringing about practical adjustments to the demands of everyday life can be measured directly by means of classroom tests.

222. Types of General Science Tests. Tests in general science may be classified either as objective achievement tests, thought and attitude scales, or as instructional tests.

[2] All tests discussed in this chapter are listed with publishers on pages 585 to 595 of *Appendix*.

[3] *Syllabus in General Science*. University of the State of New York, Bulletin No. 821. February 1, 1925, p. 7.

The former are designed to provide a basis for the evaluation of accomplishment resulting from instruction. The latter are designed to motivate learning through specific instructional units and frequent and intensive checks on progress. Each of these types of tests is discussed and illustrated in this section.

Achievement Tests. The Ruch-Popenoe General Science Test is one of the most widely used general achievement tests in this subject. It is available in two forms each consisting of two parts. Part I is made up of fifty items of general information using a seven-response multiple-choice testing technique. Part II calls for sixty-five responses based upon twenty figures each embodying a number of scientific principles. The first and last figures in Part II of the test are reproduced here as an illustration of the type of exercise used.

The content of the *Ruch-Popenoe General Science Test* was derived from an analysis of approximately twenty textbooks involving a compilation of five hundred or more facts. The percentage of each subject-matter field represented in the test is given in Table 58. The test requires forty minutes for administration. Standards are provided for grades eight and nine.

TABLE 58

RUCH-POPENOE GENERAL SCIENCE TEST
SUBJECT-MATTER EMPHASIS

Part I

Subject-Field	Per cent
Biological Sciences (Botany, Physiology, Zoology)......	30
Chemistry..	12
Physics and Mechanical Applications................	38
Earth Sciences.....................................	20

The *Powers General Science Test* consists of 200 items selected partially on the judgment of teachers experienced in general science teaching, and partially upon the basis of

an experimental try-out of some of the items. The original list of items was compiled from an examination of textbooks and courses of study in general science. The author of these tests suggests that the test results may not only reveal the pupil's level of achievement, but they may also predict his ability to do work in specialized sciences later in the high school course.

The *Dvorak General Science Scales* are available in three forms, Scales R-1, S-2, and T-2. The content was taken from general science textbooks and questions submitted by specialists in the field. The resulting 600 items were critically evaluated and then tried out experimentally. One hundred and eighty items were finally selected and incorporated into the three scales. The author states that all items of these scales satisfy one or more of the general objectives of secondary education.

The *Pieper-Beauchamp Tests* are representative of a type of standardized test designed to parallel a specific textbook. The tests are of value as pre-tests, for they furnish an inventory of the previously gained experiences of the pupil upon which later teaching can be based. They also serve as partially diagnostic measures of each pupil, thus enabling the teacher to individualize his instruction. In addition, these tests stimulate recall on the part of the pupil and link together many experiences previously unrelated in his mind. This type of test holds great promise for the future and undoubtedly is the forerunner of many more to be developed along similar lines for the several subjects of the secondary school field.

The *Van Wagenen General Science Reading Scales, Scales A and B* consist of fifteen short paragraphs on various topics of general science, followed by four to six questions based on the subject-matter of the paragraphs. Scale A emphasizes particularly the content of biology, while Scale B deals with the content of chemistry and physics. The paragraphs are divided into three groups of five each. These scales are designed principally to measure ability to comprehend science reading matter. To answer the questions on each paragraph, the pupil must abstract the meaning of the material. Some of the questions refer to information contained within the

paragraph, while others are concerned with implications of the paragraph content.

The *Downing Range of Information Test* consists of a list of fifty technical terms in the field of science, arranged in alphabetical order. The pupil has an opportunity to react twice to each item, once by indicating the extent of his acquaintance, and later by actually explaining or defining the terms.

Tests of Thought and Attitudes. The Odell Scales for Rating Pupils' Answers to Nine Types of Thought Questions in General Science are a series of scaled answers to certain thought questions in general science. The answers to the questions used in the scale are taken from the responses made by pupils and have been rated from 0 to 10. Following the answers given in the scale a criticism is made of each answer showing its good points and its deficiencies. This scale should be helpful in evaluating answers to thought questions.

Instructional Tests. The necessity of adapting instruction to individual differences in rates of learning, and difficulties revealed by test scores, has resulted in the development of a new type of instructional technique. Teaching tests such as the *Glenn-Gruenberg Instructional Tests in General Science* provide an excellent type of individualized drill on the several elements of general science instruction. There are thirty-five tests in this series, each designed to be used at definite intervals throughout the course. The purposes of these tests as stated by the authors are as follows:[4]

"1. To diagnose learning difficulties of students and study the nature of their errors separately for each unit of general sciences. . ."

"2. To provide information about student achievement on which to base instructional practices. . ."

"3. To make frequent inventories of a student's success with a reasonable expenditure of time. . ."

"4. To reveal to both teacher and student the outstanding difficulties in learning general science, as a basis for an intelligent review program or for remedial drill throughout the semester. . ."

"5. To investigate problems relating to learning in general science, and thus make a beginning in the development of the psychology of this recently organized high school subject. This is important to both teacher and research worker."

[4] Quoted with permission from Examiner's Manual for *Glenn-Gruenberg Instructional Tests in General Science.* World Book Company, 1932.

IV. Measurement in Physics

223. Objectives in Physics. The values of precise measurement in physics are well-recognized, but the accomplishment of it has been deferred through the lack of statements of objectives sufficiently specific to afford real guidance in the selection of subject matter content. The general aims of physics teaching may be stated as follows:

1. To develop a fund of knowledge of facts, phenomena, and principles of physics on the part of the student.
2. To produce ability to solve problems involving physical laws and principles.
3. To create the opportunity for the development of skill in the techniques of laboratory procedures.
4. To stimulate the development of scientific attitudes, interests, habits of inquiry, accuracy, and certain appreciations of the importance and the accomplishments of physics in contemporary life.

224. Measurable Qualities in Physics. The measurement of the factual and informational aspects of physics is not difficult to accomplish. The understanding of the relationships of facts or the generalized idea, which is one of the main objectives of good science teaching is also capable of measurement, but it is a much more difficult accomplishment than the mere testing for memory of facts. Problem-solving abilities, valuable in that they require the student to do reflective thinking, are also measurable. These two further qualities, basic to effective work in physics, also lend themselves quite readily to measurement of two types: (1) the ability to read scientific material, and (2) the mastery of underlying mathematical skills. Effectiveness in reading science material can be measured, and the results often serve as valuable means for improving the student's accomplishment in the subject. The *Van Wagenen General Science Reading Scales* are described in this chapter. A number of other reading tests suitable for this type of measurement are described in Chapter XIV. The mathematical skills which are basic to problem solving in physics can be measured by the *Kilzer-Kirby Inventory Test for the Mathematics Needed in High School Physics.*[5] This test is

[5] All tests discussed in this chapter are listed with publishers in the *Appendix.*

described very briefly in a subsequent section of this chapter.

225. **Types of Physics Tests.** In addition to the special tests mentioned in the preceding section as measuring underlying skills conditioning achievement in physics, there are two main types of physics tests, tests of achievement and instructional tests.

Achievement Tests. Only four of the more recent achievement tests in physics are described here.

The *Black and Burlingame Elementary Physics Test* is one of a series of Harvard tests. It is designed to measure the student's knowledge of physics at the end of the course. It is a six-page test of sixty items arranged in three parts. Part I contains 26 true-false items; Part II, 11 completion items; and Part III, 23 multiple-choice items. The correlation of Form A with Form B based on responses from 1741 cases is .80.

The *Columbia Research Bureau Physics Test* is a general achievement test giving emphasis to the units of physics in the proportions specified: mechanics, 16%; light, 16%; heat, 16%; sound, 8%; electricity, 32%; and miscellaneous, 12%. It consists of 144 true-false statements, including both information and reasoning questions, but with greater emphasis upon the latter. Since the test exercises are all of one type, the test is almost entirely self-administering. A special key greatly facilitates the scoring. It is intended for use as a final examination in high school physics or as an entrance test in college for placement and educational guidance. Percentile norms are provided.

The *Hughes Physics Scales* are four in number, two for measuring information and two for measuring thought. Each of the tests comprises 30 items arranged in three groups, each group being progressively more difficult. The test material was selected from the recommendations of physics teachers plus the careful analysis of textbooks. Division I of each test is intended for secondary school use, while Division II is intended for college classification purposes. These tests may be used (1) for the measurement of progress throughout a year's work in physics, (2) for diagnosis, (3) for ranking, grading and classification of pupils, and (4) for college classification. The thought

scales have a reliability of .88, and the information scales have a reliability of .85. The working time for the former scale is twenty minutes, and for the latter scale, thirty-seven minutes.

The *Stewart-Ashbaugh Physics Tests* are achievement tests designed to be given at the end of each semester. The first semester test covers mechanics and heat while the second semester test covers electricity, sound, and light. The tests measure a wide range of abilities and facts. One of the significant abilities emphasized is that of the association of a device or process with the physical principle which it involves. The tests are in two forms. The working time for each test is thirty-five minutes. The reliabilities of the first semester and the second semester tests are respectively .83 and .89.

Instructional Tests. Instructional tests open many possibilities for the promotion of diagnostic and remedial work in the sciences. Tests of this type in mathematics and in general science have been described previously. Two of the best representatives of this type of test in the field of physics are presented here mainly for the purpose of showing their possibilities for improving instructional effectiveness.

The *Glenn-Obourne Instructional Tests in Physics* comprise twenty-five unit tests totaling 973 separate items covering the entire field of physics. Each test is to be used at the completion of a unit of subject matter to reveal any weaknesses in the unit completed, and to indicate the pupil's readiness to undertake the next unit. These tests are most useful in a beginning course in high school physics and are especially valuable for diagnostic work, motivation and marking. Reliable norms are available. The reliabilities of the twenty-five tests range from .70 to .88.

The *Michigan Instructional Tests in Physics* consist of : (1) an initial test which has for its purpose the determination of the fitness of the student to pursue the course in high school physics, and the analysis of the strengths and weaknesses of the prospective student ; (2) a series of instructional tests, twenty-two in number, each test covering fairly completely a certain phase of high school physics, the tests

to be applied for the purpose of making a continual check upon the pupil's achievements during the process of instruction; and (3) a final examination to be applied at the end of a semester or year to obtain a final measurement of the quality of the student's work in the semester or year.

The *Initial Test*, which is intended to be applied previous to instruction in physics, determines what general information the student has in physics, how well he can handle certain types of mathematical problems involved in physics, and how effectively he can read and understand physics material.

The *Instructional Tests* in physics cover the various phases of physics. The following types of test items are included: (a) mathematical problems in physics which demand a knowledge of physics subject matter, (b) completion or multiple choice test items involving a knowledge of the laws and principles of physics, (c) problems involving the mathematical statement of laws and principles of physics, and (d) problems involving the interpretation of figures in terms of physical laws and principles.

The *Final Examination* is constructed from a random sampling of the test items included in the various instructional tests. It is made in such a way that under each of the various major phases of physics subject matter (such as mechanics, heat, electricity and magnetism, sound, and light), there appears a series of test items of known value selected at random from the various instructional tests covering the same field of subject matter.

226. Miscellaneous Physics Tests. The *Kirkpatrick-Greene Handbook of Objective Test Exercises in Physics* consists of a comprehensive series of test items covering the basic concepts of the subject. The important items were determined by checking their appearance in seven of the more recent texts in high school physics. Several different objective test forms are used for each of the important concepts making it possible for the teacher to vary the testing technique for quiz and examination purposes. The exercises include thought-provoking questions, simple fact, questions and problems. They are of value to any teacher of the subject since they are adapted for use in connection with

any textbook or course of study. The teacher himself validates the content of his test or examination by selecting from the handbook the items which are pertinent to the unit of subject matter he has been teaching.

Recent investigations show that in some schools approximately one-fourth of the pupils who take physics fail to make a passing grade in the subject; that physics pupils use their mathematics poorly; that teachers omit a large percentage of the problems given by the textbooks; and that mathematical ability is one of the best criteria for predicting success in high school physics. In view of these facts the *Kilzer-Kirby Inventory Test for the Mathematics Needed in High School Physics* was constructed. The purposes of this test are: (1) to point out to mathematics teachers those items in their subjects which are useful in high school physics, and to provide a means for testing the pupils on these items; (2) to assist the administration in guiding pupils in their choice of high school physics; and (3) to provide an inventory test which may be given by the physics teacher during the first week of the course.

The *Iowa Placement Examinations in Physics* are designed particularly for use with college students in discovering special aptitudes as well as in determining the effect of preliminary training in the field. The aptitude series is designed to measure the following factors which appear to contribute to subsequent success in a subject:

Part 1. Simple arithmetic of physics
Part 2. Ability to secure precise data from physics paragraphs
Part 3. Logical relations
Part 4. Interest in physics

The training series measures previous training in the subject which is considered an essential preliminary to the work at hand and consists of:

Part 1. Fundamental knowledge and principles
Part 2. Fundamental knowledge and principles
Part 3. Equation and law
Part 4. Fundamental physical problems

Reliable norms are available for each part of this examination.

Lapp's Iowa Achievement Examinations in College Physics are objective examinations in college physics which parallel a widely used textbook in college physics and are so arranged as to take the place of the traditional written examination covering this subject matter. The series of eight examinations covers the complete year's work. They are now available in two equal forms, Series A and B. Tests 1, 2 and 3 of each series are the mid-semester tests for the first half-year. Test 4 is the final for the first semester. Tests 5, 6, 7 are mid-semester tests for the second half-year, and Test 8 is the second semester final. Standards are available for each examination.

V. Measurement in Chemistry

227. **Objectives of Chemistry.** While chemistry is one of the older sciences the general lack of agreement in its objectives is just as apparent as in other fields. The commonly accepted objectives are extremely long-range, and as such are not particularly helpful in the selection and organization of curricular content which must precede measurement in the subject. The committee responsible for the preparation of *A Program for Teaching Science*[6] suggested three objectives of chemistry teaching as follows:

1. Pupils in high school chemistry should develop better understanding of those fundamental concepts, major ideas, laws, or principles of chemistry that will enable them better to interpret natural phenomena, common applications of chemical principles, and industrial applications and uses of the principles of chemistry.

2. They should learn to use the processes of reflective thinking, problem solving and techniques of study that are best adapted to the solution of problems within the field of chemistry, especially those which most often present themselves in daily life.

3. They should develop those attitudes toward the facts and principles of chemistry and toward the methods of

[6] Powers, S. R. (Chairman) et al., *A Program for Teaching Science.* Thirty-first Yearbook, Part I, of the National Society for the Study of Education. Bloomington, Illinois: Public School Publishing Company, 1932.

investigation employed in the field that will serve as guides in their use of chemical facts and principles and methods of problem solving.

The more specific aims of the instruction in the subject may be stated as follows:

1. To develop a functional knowledge of the scientific facts, phenomena, principles and laws of chemistry as they apply to the improvement of human welfare.
2. To develop the ability and willingness to solve problems using these principles.
3. To develop dynamic scientific interests and attitudes.
4. To develop effective laboratory procedures.
5. To prepare adequately for direct vocational needs or for later advanced work in chemistry.
6. To enable the child better to understand and adjust himself to his environment.

228. **Measurable Qualities in Chemistry.** Many of the implied outcomes of instruction in chemistry cannot be measured accurately and objectively. It is possible, however, to measure with reasonable accuracy such elements as (1) factual matter or informational items of the course, (2) specific techniques, such as determining valence relationships and balancing equations, (3) ability to solve chemical problems, (4) laboratory procedures, and (5) certain aspects of scientific attitudes.

229. **Types of Chemistry Tests.** A few representative tests in chemistry are discussed briefly under the major classifications of achievement tests and instructional tests.

Achievement Tests. The *Columbia Research Bureau Chemistry Test* contains over 250 items covering the materials common to the most popular high school and introductory college textbooks in chemistry and also the college entrance requirements. About 50% of the test is devoted to general information and definitions, about 29% to formulas and equations, and about 21% to problems. Each of these three aspects of chemistry is covered in a separate part. Reliable norms are provided. The time required for administering the tests is two hours.

The *Powers General Chemistry Test* is designed to measure the accomplishment of students in a general course

in chemistry. It consists of two parts: Part I, a test of range of information consisting of 30 questions, and Part II, a test of ability to do tasks in chemistry consisting of 37 problems. Two forms, A and B, are available and each is provided with reliable age, grade and percentile norms. The working time is thirty-five minutes.

The *Gerry Test of High School Chemistry* is based upon 96 questions taken from the College Board Examinations and checked against four popularly used textbooks. The test is available in two forms, A and B, each containing 25 informational items. The test is designed to occupy a single class period. Because of its brevity, the reliability of the test is limited.

Rich Chemistry Tests for High Schools are available in two forms, *Gamma* and *Epsilon.* These tests, consisting of 25 multiple-choice questions in each form, are designed to measure achievement in general chemistry. T-scores are used and are comparable from one test to another. The attainment is measured in terms of: (a) ability to think; (b) information; (c) ability to solve numerical problems; and (d) habits and knowledges acquired in the laboratory. The tests are arranged to call out these abilities in cycles of six questions each. It is doubtful whether in the working time of 25 minutes provided for the test, the four purposes of the tests can actually be accomplished.

The *Iowa Placement Examinations, Chemistry Aptitude* and *Chemistry Training* in *Forms X* and *Y* are designed particularly for use with college students in discovering special aptitudes for the study of chemistry as well as in determining the effect of preliminary training in the field. The training series assumes one high school year or one college semester of training. The aptitude series requires no special training in the subject, but is designed to measure those factors which contribute to subsequent success in a subject. It consists of the following points:

Part 1. Ability to comprehend and apply rules
Part 2. Ability to secure correct ideas from English textbook material
Part 3. Reading comprehension
Part 4. The fundamental chemical problems

This aptitude test is the best representative of this type of test. It is held to be highly suitable for the following purposes:

1. To afford a basis for prediction of the character of work that each student will do in college.
2. To aid in selecting and admitting students.
3. To serve as an entrance examination in lieu of the more time-honored essay type of content examination.
4. To section classes for instructional purposes on the basis of mental ability.
5. To assist in deciding how much work a student can carry.
6. To deal more effectively with students who are not well oriented in their college work.
7. To give scientific aid in vocational guidance and placement.
8. To enable comparative studies of intellectual levels as between classes, colleges and college years.
9. To furnish a basis for the diagnosis of class weaknesses.
10. To enable high schools and preparatory schools to survey their seniors or graduates in order to determine probable fitness for collegiate studies.

Reliable norms for prognostic purposes are available.

Instructional Tests. The *Glenn-Welton Instructional Tests in Chemistry* are quite similar in general arrangement to the *Glenn-Gruenberg Instructional Tests in General Science* and the *Glenn-Obourne Instructional Tests in Physics* which are described briefly on pages 426 and 429 in this chapter. The *Glenn-Welton Instructional Tests* comprise a 76 page booklet containing 36 objective and standardized unit tests covering 1447 items. The reliabilities of these tests range from .70 to .88 with an average of .79. Percentile norms for each of the tests are provided. A record of progress kept by each individual student serves to motivate the work.

VI. Measurement in Biology

230. Instructional Aims in Biology. According to the yearbook committee responsible for Part I of the Thirty-first Yearbook of the National Society for the Study of Education[7] "the course in biology, stated in terms of its outcomes, will consist of (1) an understanding of the important prin-

[7] *Ibid.*

ciples of biology so that the pupil may apply them to the solution of such problems of a biological nature as will arise in his life; (2) an appreciation of some of the scientific attitudes exemplified in the work of such great biologists as Vesalius, Malpighi, Pasteur, and a sense of the lawfulness of nature and of every man's obligation to obey such laws; and (3) a reasonable degree of skill in the use of the scientific method of thinking on matters biological."

From this general statement of outcomes the following somewhat more specific aims may be summarized as a more definite guide to curricular content and measurement procedures:

1. To provide the student with a first hand knowledge of biological facts, phenomena and principles of importance to him as a human being.
2. To give him a functional knowledge of the principles of hygiene and sanitation and their relation to human welfare.
3. To develop an ability and willingness to attack problems using biological methods and principles.
4. To establish an appreciation of the importance and value of a knowledge of biology, hygiene and sanitation in contemporary life.

231. **Types of Biology Tests.** The majority of objective tests in the field of biology deal almost entirely with the measurement of facts, information, and principles. The less tangible outcomes, as interests and appreciations, have consistently avoided identification and measurement.

Achievement Tests. The *Ruch-Cossmann Biology Test* is one of the widely used tests in the field of general biology. The validation of the items in the tests is largely based upon the analysis and rating by competent judges of over two thousand final examination questions collected from all parts of the country. The materials represent the combined judgments of teachers and authorities on biology teaching. The test is available in two forms, A and B, and consists of five separate tests containing the specified numbers of items:

1. General biological information .... 40 items
2. Best answer statements .... 18 items
3. Identification of structures .... 15 items
4. Laws of Mendelian inheritance .... 4 items
5. Completion exercises .. 35 items

The time required for the entire test of 112 items is 38 minutes. The test is designed for use in high school and for testing entering college freshmen. Percentile norms are given.

The *Oakes and Powers Test of General Biology* is available in two forms, each requiring about 35 minutes working time. It is designed to measure range of information and knowledge of principles in biology and the ability to apply them to concrete situations. The 100 items comprising each form were selected from a list of 580, which had been checked against the New York Regent's examinations, other tests, and a number of outstanding textbooks.

The *Cooprider Information Exercises in Biology* have as their function the discovery of general biological knowledge. Approximately 1200 items selected from widely used textbooks, have been arranged in 25 tests. These items have been arranged in six sets of exercises as follows: (1) completion sentences; (2) recognition of true and false statements; (3) test of information or knowledge of biological terms; (4) test of reasoning ability; (5) classification of terms; and (6) logical selection. The tests have been helpful in furnishing: (1) a reliable and objective score to supplement the judgment of the teacher in making promotions; (2) a quick way of arriving at the probable biological information of classes at the beginning of the year and at mid-year; and (3) a check on the progress made during the course. Percentile norms are available.

The *Michigan Botany Test* consists of one form of four distinct parts requiring a total working time of 25 minutes. The purpose of the test is to determine achievement after one year's study of botany in the high school. Test I is of the true-false type; Test II of the completion type; Test III of the matching type; and Test IV, the reasoning type. Tests I, II and III are each composed of 20 exercises and Test IV is made up of 10 problems. The reliability of the test is reported as .87.

Instructional Tests. Tests of this type are primarily different from other standard tests in that they consist of series of tests of attainment each covering a different unit of work and planned to be given at frequent intervals through-

out the course. They have the effect of quite definitely directing instructional emphasis.

The *Blaisdell Instructional Tests in Biology* consist of a series of 25 comprehensive tests, each covering a unit of work of the customary high school course in elementary biology. Given at ten-day intervals throughout the course, they afford a highly reliable basis for adapting instruction to class and individual needs. Tests of this type give the teacher information as to the extent of the student's mastery of each unit before passing to the next; they provide definite standards of achievement; and they save time in making up weekly or monthly tests.

VII. Miscellaneous Types of Science Tests

232. **Informal Objective Tests in Science.** Because of a lack of agreement on the minimal essentials for each of the science subjects and the wide variation in the sciences taught in the different high schools, standardized tests often fail to meet the teacher's needs for testing. Frequently the instructional emphasis in certain of the standardized tests does not conform to the local course of study. This results in tests which are definitely lacking in validity for the individual school or class. In such a situation the use of informal objective exercises based upon subject-matter content definitely emphasized by the instructor is a logical and proper solution. The characteristics, construction, use and values of informal objective examinations are discussed at length in Chapter IV.

The labor of constructing informal objective examinations could be greatly reduced through the more extensive preparation and use of handbooks of objective exercises in the special subjects. Such handbooks would clarify the minimal essentials of the course. Furthermore, they would tend to improve the general quality of informal testing by recognizing the lack of time and skill on the part of the busy teacher for preparing valid and discriminating objective exercises. Such handbooks would list all of the important phases of the subject in different types of objective forms. Thought questions as well as simple fact questions and problems could be

included. Something of this same purpose is served by certain of the instructional tests, but in these cases the instructional emphasis of the course is completely dominated by the point of view represented by the maker of the test. The handbooks would make it possible for the teacher to prepare a valid examination for the class, since he would be able to select those items which have received instructional emphasis in the past. The reliability of the test would be controlled by the number of items and the length of the tests used. Objectivity would be assured by the form of the exercises themselves.

233. **American Council Co-operative Tests.** These tests are prepared by the Co-operative Test Service under a subvention from the General Education Board. The purpose is to construct ten or more comparable forms of examinations in the fundamental subject matter of junior and senior high school levels, and to make them available to the colleges and schools — one form each year, at the lowest possible cost. Colleges and schools may inaugurate the use of these tests with the assurance that a new comparable form of each will be made available each year about April 15th. Some of the advantages of such a series of comparable examinations are that —

a. Individual growth in defined types of achievement can be measured year after year, thus making feasible types of educational guidance and of educational research which would be difficult or impossible with unrelated examinations or with "standardized" tests which exist in two or three comparable forms.
b. Each college, school, or school system that uses these tests may set up its own standards for admission, placement, promotion, certification, and gradation, and maintain those standards uniformly from year to year, and at the same time maintain transfer and advanced standing relations with other institutions on the basis of comparable achievement test measurements.
c. The results of subjective and other local examinations can be made closely comparable from year to year by using the results of the co-operative tests as a common denominator, thus taking advantage of the best features and minimizing the weaknesses of both types of examinations.

These tests are now available in high school biology, chemistry, general science, and physics. Tests for college

use are also announced in astronomy, biology, general science, geology, physics, and zoology. This series of tests is discussed briefly on page 554 in Chapter XXII.

234. **Iowa Every-Pupil Tests in Science.** Among the nonstandardized tests which have received wide attention are the *Iowa Every-Pupil Academic Tests.* These tests are prepared annually in connection with a state-wide academic contest in the secondary schools. In addition to their motivating values and competitive features, these tests provide a unique source of important information relative to curricular content and emphasis. Most standardized achievement tests are intended for nation-wide use. This very attempt to make them generally valid partially destroys their specific validity. The idea back of the development of the *Iowa Every-Pupil Tests* is that the best type of test for the schools of a state is the test which is designed specifically to fit the local instructional situation.

The content and the organization of these Iowa Every-Pupil Tests varies from year to year so that it is impossible to describe them completely. In the 1935 series, tests in biology, general science, and physics were included. The summary given in Table 73, page 554, indicates something of the nature and content of these science tests. The complete series of tests is described in more detail in Chapter XXII.

VIII. Diagnostic and Remedial Work in the Sciences

235. **Diagnosis and Remedy.** Teachers of the secondary school sciences find that the science tests now available are inadequate for the diagnosis of pupil difficulties. While certain of the available tests may show students to be deficient in some specific phases of science information, the majority of the tests do not point out the cause of such deficiencies. Practically all that can be done by way of exact diagnosis is in connection with certain skills which appear to underlie the study of science. In general, the causes of failure in the science courses will be found to be due to one or more of the following factors:

1. Mental ability.
2. Ability to read and comprehend science material.
3. Knowledge of technical vocabulary.
4. Arithmetical ability (especially in physics).
5. Algebraic ability (in physics and chemistry).
6. Lack of interest and application.

Information relative to the student's mental ability may be secured with reasonable reliability through the use of one of the group tests of mental ability discussed in Chapter XI. This information should be supplemented by data from silent reading tests. Reading of science material seems to involve the comprehension of a language peculiar to the subject, and is apt to be difficult. Diagnosis of reading abilities of the work-study type, accompanied by remedial instruction designed to overcome the weaknesses revealed, is one of the prerequisites to satisfactory progress in the study of the sciences. Another approach to the analysis of a student's poor work in science may be made through the use of science vocabulary tests such as the *Pressey Technical Vocabulary Lists* or similar ones prepared by the instructor. Survey tests in arithmetic and inventory tests such as the *Kilzer-Kirby Tests* should provide useful information on the causes of student difficulties in subjects in which achievement depends upon these specific abilities.

If the student's deficiency consists in poor reading ability, an inadequate knowledge of the technical vocabulary of the subject, a limited ability to apply his mathematical skills, and not in low mental ability, the obvious task of the teacher is to administer drills designed to improve these aspects of the study procedure. Inventory, diagnostic, and instructional tests are of considerable value for motivating purposes, and in many cases may solve the problem of failure due to lack of interest and application. Laboratory work may call for many new abilities and techniques. It is important, therefore, that guidance be given the student to insure the early discovery of his weaknesses, and to provide for him the opportunities to develop sufficient mastery over them to insure satisfactory accomplishment in this phase of his school work.

Discussion Exercises

1. Why is objective measurement in the sciences not highly developed?
2. Enumerate and evaluate the chief general aims of education which presumably are met by science instruction.
3. In your opinion is the need for a unified course in high-school science any less serious than it is in the social studies?
4. In what ways are the aims of teaching high school physics distinct from those for the teaching of general science?
5. What are the four most important measurable qualities in science achievement?
6. Examine the possibilities of measuring the acquisitive skills in science and specify a number of techniques in each. Would such a list parallel the types found in the social studies?
7. Make a list of specific outcomes of instruction in general science.
8. Outline a series of corrective lessons for a class which is having difficulty with the work in general science, or some other science course.
9. In your opinion what is the relation of *facts* to *problem-solving* in the sciences?
10. Examine the available tests in the fields of biology, chemistry, general science, or physics and select one test for each giving your reasons for your choice.

Selected References

Almack, John C., and Lang, A. R., *The Beginning Teacher*, Chapter XIX (Teaching the Natural Sciences). Boston: Houghton Mifflin Company, 1928.

Beauchamp, Wilbur L., *Instruction in Science.* Monograph 22, National Survey of Secondary Education, Bulletin No. 17. Washington, D. C.: United States Department of Interior, Office of Education, 1932.

Brownell, H., and Wade, F. C., *The Teaching of Science and the Science Teacher.* New York: The Century Company, 1925.

Burgess, T. O., *A Psychological Analysis of Abilities in High School Physics.* University of Iowa Studies in Education, Vol. 3, No. 6. Iowa City: University of Iowa, 1926.

Craig, G. S., *Techniques Used in Developing a Course of Study in Science for the Horace Mann Elementary School.* New York: Teachers College, Columbia University, 1927.

Cunningham, H. A., "Character and Value of Existing Tests, for

Pupils and Teachers in General Science," *General Science Quarterly*, Vol. 13:61-69, January, 1929.

Curtis, F. D., *A Digest of Investigations in the Teaching of Science in the Elementary and Secondary Schools.* Philadelphia: P. Blakiston's Sons and Company, 1927.

Department of Superintendence, *Fourth Yearbook. The Nation at Work on the Public School Curriculum,* Chapter IV, "Elementary Science and Nature Study." National Education Association, Washington, D. C., 1926.

Department of Superintendence, *Fifth Yearbook. The Junior High School Curriculum.* Washington, D. C.: National Education Association, 1927.

Department of Superintendence, *Sixth Yearbook. Development of the High School Curriculum.* Washington, D. C.: National Education Association, 1928.

Downing, E. R. (Chr.) "Teaching Units in Biology—an Investigation." *North Central Association Quarterly,* Vol. 5: 453-470, March, 1931.

Freeland, G. E., Adams, R. M., and Hall, F. H., *Teaching in the Intermediate Grades,* Chapter IX. "Nature Study and Science." Boston: Houghton Mifflin Company, 1927.

Hammond, H. P., and Stoddard, G. D., *A Study of Placement Examinations.* University of Iowa Studies in Education, Vol. 4, No. 7. Iowa City: University of Iowa, 1928.

Hurd, A. W., "Reorganization in Physics," *North Central Association Quarterly,* Vol. 4: 277-93, September, 1929.

Kilzer, L. R., "The Mathematics Needed in High School Physics," *Science Education,* Vol. 14: 335-44, November, 1929.

Kirkpatrick, J. E., *The Motivating Effect of a Specific Type of Testing Program.* University of Iowa Studies in Education, Vol. 9, No. 4. Iowa City: University of Iowa, 1934.

Odell, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

Peters, C. J., and Watkins, R. K., *Objective Tests for High School Physics.* Columbia, Missouri: C. J. Peters and R. K. Watkins, 1926.

Powers, S. R. (Chairman) et al., *A Program for Teaching Science.* Thirty-first Yearbook, Part I, National Society for the Study of Education. Bloomington, Illinois: Public School Publishing Company, 1932.

Powers, S. R., "The Vocabularies of High School Science Textbooks," *Teachers College Record,* Vol. 26: 368-83, May, 1925.

Rich, S. G., "Achievement of Pupils in Chemistry," *School Science and Mathematics*, Vol. 25 : 145-149, February, 1925.

Rich, S. G., "The Available Tests for Results of Teaching the Sciences," *School Science and Mathematics*, Vol. 26 : 845-52, November, 1926.

Richards, O. W., "Test Construction in Less Standardized Subjects Illustrated by the Richards Biology Test," *School Science and Mathematics*, Vol. 27 : 22-27, January, 1927.

Smith, H. L., Wright, W. W., *Tests and Measurements*. New York : Silver, Burdett and Company, 1928.

Symonds, P. M., *Measurement in Secondary Education*. New York : The Macmillan Company, 1927.

CHAPTER XVIII

MEASUREMENT IN THE FOREIGN LANGUAGES

THE purpose of this chapter is the summary of the following points involved in measurement in the foreign languages:

a. Importance of instruction in foreign language.
b. Objectives of foreign language instruction.
c. Identification of measurable factors.
d. Measurement, prognosis and remedy in Latin.
e. Measurement, prognosis and remedy in modern languages.

I. EDUCATIONAL IMPORTANCE OF THE FOREIGN LANGUAGES

236. **Shifting Objectives of Instruction.** The study of the foreign languages has held its place in the secondary schools and colleges because it has seemed to provide an effective means of transmitting racial culture and of developing historical perspective. In the past the classical languages have been defended because of their disciplinary values. Today their chief defense lies in the attention they focus on the English language itself. Teachers of foreign language are quite generally agreed that the two cultural objectives, namely, "(1) knowledge of the foreign country and its people, and (2) increased knowledge of English words, English grammar, and relationships between the foreign language and English,"[1] are the major outcomes of instruction in the languages.

In the modern languages, two quite definite trends are pointed out in the more specific objectives of classroom activi-

[1] Eddy, Helen M., *Instruction in Foreign Language.* Monograph 24, National Survey of Secondary Education, Bulletin No. 17. Washington, D. C.: United States Department of Interior, Office of Education, 1932.

ties. One holds to the traditional aims of developing the ability of the student: (1) to read, (2) to write, (3) to speak, and (4) to understand the language. Another more liberal point of view holds to the two cultural objectives named above plus the "progressive development (a) of ability to read books, newspapers, and magazines in the modern language within the scope of the student's interests and intellectual powers; (b) of such knowledge of the grammar of the language as is demonstrated to be necessary for reading with comprehension; (c) of the ability to pronounce correctly, to understand, and to use the language orally within the limits of class materials."[2]

Cole[3] suggests the following immediate aims as guides for the teacher in shaping instruction since the commonly accepted ultimate objectives are too general for practical application in classroom work.

1. The ability to read with increasing ease material in the foreign languages within the scope of the student's interests and intellectual powers.
2. A sufficient knowledge of the structure of the language to read understandingly.
3. The ability to pronounce correctly, to understand, and to use the language orally within the limits of materials used in class.
4. The ability to write the language for class purposes with a relatively higher attainment of fluency in vocabulary and in idiom than of grammatical correctness.
5. The acquisition of some knowledge of the country whose language is being studied, and of a special interest in the life and characteristics of its people.
6. The acquisition of a knowledge of the general character and history of the modern language and of its relationship with English.

Of these objectives, 1, 2 and 5 are probably the most important, particularly if the modern language course is organized on a two-year basis.

The objectives of instruction in Latin naturally are somewhat different. The most recent courses of study accept the statement of aims, both ultimate and immediate, for instruc-

[2] Coleman, Algernon. *The Teaching of Modern Foreign Languages*. Publications of the American and Canadian Committees on Modern Languages, Vol. XII. New York: The Macmillan Company, 1929.

[3] Cole, Robert D., *Modern Foreign Languages and Their Teaching*. New York: D. Appleton and Company, 1931.

tion in Latin as recommended by the *Classical Investigation.*[4] The immediate objective is the progressive development of the power to read and understand Latin. This involves an increasing mastery of vocabulary, forms, and syntax. The ultimate objectives, according to this report, include:

a. Increased understanding of those elements in English which are related to Latin
b. Increased ability to read speak, and write English
c. Increased ability to learn other foreign languages
d. Development of correct mental habits
e. Development of an historical and cultural background
f. Development of right attitudes toward social situations
g. Development of literary appreciation
h. An elementary knowledge of the simpler general principles of language structure
i. Improvement of the literary quality of pupil's written English

II. Measurement of Latin

Measurement in Latin has thus far been restricted mainly to the immediate objectives of the subject. Effective measurement in Latin, as in every other field of learning, is dependent upon the exact identification of the specific elements which comprise the desired outcomes of instruction. To attempt to measure all of the objectives of instruction in Latin would lead to confusion and almost certain failure. Many of these objectives emphasize cultural and disciplinary values which at present are not measurable.

237. **Possibilities of Measurement.** The analysis of achievement in Latin indicates that the accomplishment of the immediate objectives in the subject is dependent upon a variety of specific abilities many of which are measurable. Among these are such factors as the acquisition of vocabulary, and the oral and silent reading skills. The knowledges and skills listed here are suggestions of the types of measurable elements in Latin instruction:

1. Acquisition of vocabulary
   a. English into Latin
   b. Latin into English
2. Knowledge of vocabulary forms

[4] American Classical League, *The Classical Investigation, Part I, General Report.* Princeton, New Jersey: Princeton University Press, 1924.

3. Syllabication and accentuation skills
4. Mastery of verb forms
   a. Latin into English
   b. English into Latin
5. Knowledge of noun and adjective forms
6. Knowledge of noun and pronoun meaning and case use
7. Skill in syntactical analysis of English passages
8. Knowledge of derivations of word forms
9. Comprehension of material read in Latin
10. Speed and accuracy in translation
    a. English into Latin
    b. Latin into English
11. Spelling of required forms
12. Understanding of Latin words, phrases, abbreviations, and quotations appearing in English
13. Knowledge of historical perspective and general cultural background, resulting from study of Latin
14. Understanding and appreciation of allusions to the mythology, traditions, and history of the Greeks and Romans.

238. **Achievement Tests in Latin.**[5] Numerous Latin tests designed to measure one or more of these elements involved in the accomplishment of the more immediate objectives of instruction in Latin are now available. By far the majority of these tests are intended to measure general achievement rather than specific factors underlying achievement. Many parallel quite closely the testing techniques developed in the field of English. Relatively little has been accomplished in the development of highly analytical or diagnostic instruments in this field. A few typical achievement tests in Latin are described in this section.

The *Godsey Latin Composition Test* is designed to gauge objectively the accuracy of composition work, and at the same time to determine the correlation of sentence-writing skill with knowledge of grammar. The test was administered to many thousands of students in the nation-wide survey conducted by the American Classical League. The test comprises thirty-three exercises. Each exercise consists of an English sentence and its translation into Latin. In the translation one word is given in four forms. The student is directed to draw a circle around the correct form. Numbers, placed at the end of each Latin sentence, refer to gram-

[5] All tests mentioned in this chapter are listed with publishers in the *Appendix*, pages 585-595.

matical rules which are printed at the bottom of the test folder. The student is directed to draw a circle around the number of the rule which applies to the correct form. A sample exercise is given below:

1. They were hurling weapons at the wall.
Tela in murum (coniciebantur, coniciebant, coniciunt, conicierent)
1 .8... .13........14

There are two equivalent forms of the test, A and B. The time for administering the test is about 35 minutes. Norms are based upon approximately 20,000 cases. The test may be used in any year of high school Latin.

Henmon's Latin Tests are the result of seven years' experimentation. The tests are prepared in two series, one for measuring the students' acquaintance with Latin vocabulary and the other for measuring his ability to translate Latin sentences into English. Four vocabulary tests of fifty words each are constructed so that they are essentially duplicate forms. In each test the words are arranged in ascending order of difficulty. For measuring ability of students to translate Latin into English, sentences are used in which only words occurring in the fundamental vocabulary list appear. The difficulty of these sentences has been determined, and they are arranged in ascending order of difficulty. The student is asked to write his translation in English.

The *Hutchinson Latin Grammar Scales* measure one of the immediate objectives of Latin instruction in the secondary school, i.e., the functional knowlege of Latin syntax. The tests are available in two forms, Scale A and Scale B. Each scale consists of 35 multiple-choice exercises, of which 25 are on nouns, pronouns, and adjectives, and 10 on verbs. The test items were chosen on the basis of (1) presence in ten commonly used first-year textbooks and (2) frequency in high-school Latin.

Standards are provided for Scales A and B for students who have had one or more semesters of Latin. The two forms have a coefficient of correlation of .854 ± .008. The working time for the test is twenty-five minutes.

The *Lohr-Latshaw Latin Form Test* is designed to measure the knowledge of forms. The test consists of 35

Latin forms of nouns, pronouns, and verbs which are selected to be of nearly equal difficulty. The student is required to indicate his information about each of the forms under the proper headings. This test is valuable in revealing lack of sufficient strength in one of the three parts of speech — nouns, pronouns, and verbs, to permit the student to carry on his study of Latin effectively. The working time for the test is twenty minutes.

The *New York Latin Achievement Test* is intended especially for a reliable check on the strong and weak points of students in each phase of the subject. It is comprehensive and dependable. It tests vocabulary, syllabification and accent, forms of parts of speech, derivation, translation, adjective and adverb comparison. The tests were perfected after experimental use in two Latin surveys conducted in New York state. They include *Test* 1 for the first half year of Latin and *Test* 2 for the second half year of Latin. There are two forms of each test. The working time for the test is eighty minutes.

Powers' Diagnostic Latin Test measures the extent to which students in first-year Latin have mastered the vocabulary, word forms, and skills in translation necessary for a good foundation in the subject. The test is designed not only to survey the student's knowledge as a whole, but also to give evidence of major points of weakness in his preparation.

Two forms of the test are available, Form 1 and Form 2. Each form consists of seven distinct parts which may be given all at one time or at different times after drill upon the particular points or phases of Latin instruction. The seven parts of the test are:

1. English-Latin translation
2. Nouns and adjectives
3. Verbs
4. Vocabulary
5. Comprehension and Syntax, Part A
6. Comprehension and Syntax, Part B
7. Derivatives

Although the test is primarily intended for first-year students, it may be used to excellent advantage in canvassing

the basal preparation of a second, third, or fourth year class at the beginning of a term. The working time is two forty-five minute periods. The reliability of the entire test is .96. On the various parts, the reliability range is from .80 to .90.

The *Pressey Test in Latin Syntax* is of the multiple-choice type, consisting of thirty-two English sentences, each of which is translated into four different Latin sentences only one of which is correct. The student is expected to underscore the Latin sentence which he considers correct. The test measures the student's familiarity with Latin nouns, pronouns, and adjectives. The exercises are arranged in ascending order of difficulty. The student is allowed twenty minutes for the test. Norms are provided for students having had two to eight semesters of Latin. A sample exercise follows:

Caesar drew up a triple line of battle.
Caesar aciem triples instruxit.
Caesar aciem triplicam intruxit.
Caesar aciem triplicem intruxit.
Caesar aciem triplici instruxit.

The *Stevenson-Coxe Latin Derivative Test* is designed to check the ability of pupils to recognize the Latin in English words. The students are required to look at each English word given and think of the Latin word or words from which the English syllables may be derived. Then they are expected to find in a column of Latin words the one which forms an important part of the English word. Norms are provided for students having had one, two, three, and four semesters of Latin. The time limit for this test is fifteen minutes.

Stevenson's Latin Vocabulary Test consists of sixty Latin words arranged in columnar form. In parallel columns, 120 English meanings are presented, each of which is numbered. The pupil is required to write before each Latin word the number giving its English meaning in the opposite column.

Standards are provided for grades 7 to 10, for students having had two, four, or six semesters of Latin. The reliability correlation of the Forms 1 and 2 is .86. The working time for the test is set at fifteen minutes.

The *Tyler-Pressey Test in Latin Verb Forms* uses the multiple-choice type of testing technique. It consists of thirty-two Latin verb forms, each of which is translated into four English verb phrases, only one of which is correct. Norms are provided for pupils having had two to eight semesters of Latin. The time limit for the test is 15 minutes. A sample exercise is given below:

1. Movemus

We shall move. . .	Let us move. . .
We are moved. . .	We are moving. . .

The *Ullman-Kirby Latin Comprehension Test* is planned to measure the student's ability to comprehend the thought in the Latin passage. The content of this test consists of passages similar to the texts read in the secondary school. These passages are arranged somewhat like the *Thorndike-McCall Silent Reading Scales.* Each of the ten Latin passages or paragraphs is progressively more difficult than the preceding one. Answers to the questions based on each passage are to be given in English and are designed to reveal the degree to which the student is able to comprehend what he has read. The test is prepared in two equal forms, which may be used in any year of the Latin course.

Exercises from the first and the final paragraphs of Form 1 of this test are reproduced here as an illustration of the content of the test.

Ullman-Kirby Latin Comprehension Test [6]

PARAGRAPH I

Read this and then write the answers. Read it again if you need to.

Puer Rōmānus erat quī Mārcus appellābātur. Idem nōmen pater habēbat quod fīlius. Pater puerī magnam domum Rōmae habēbat. In hāc urbe puer et māter habitābant. Pater mīles in Galliā erat, ubi multōs annōs pugnāverat. Etiam puer mīles in exercitū esse volēbat.

[6] Ullman, B. L., Kirby, T. J., *Ullman-Kirby Latin Comprehension Test.* Reproduced with the permission of the publishers, Bureau of Educational Research and Service, University of Iowa, Iowa City.

1. What was the name of the boy's father? . . . .
2. Which members of the family were in Rome?
3. What did the boy want to be?
4. What had his father done?

PARAGRAPH X

Read this and then write the answers. Read it again if you need to.

Nūntius exanimī suspēnsus pectora cursū
Aepytus ad rēgem, portae statiōne relictā,
tendit et haec trepidō vix intellēctus anhēlat:
"Linque piōs cultūs intempestīvaque, rēctor,
sacra deum: frāter mūrīs circum omnibus īnstat
nōmine tē crēbrō, tē sōlum in proelia poscēns."

31. Why was it hard for the king to understand what the messenger said? . . . .
32. What does the messenger urge the king to stop doing?
33. What did the king's brother demand?

The *White Latin Test* is designed to measure growth in general knowledge of Latin. It consists of two parts, a vocabulary test and a translation test. Part I, the vocabulary test, comprises 100 words selected according to frequency of occurrence in the works of authors that are read for college entrance. Part II is a test in the translation of sentences. It contains twenty sentences grouped according to increasing difficulty of syntax. Alternative answers are given in both parts so that the test can be scored rapidly. Scores on this test may be used as a basis for dividing students into sections according to ability, and as an aid in determining promotions. Semester norms are available for first, second, third, and fourth year Latin. The working time for the test is 35 minutes.

239. **Prognosis in Latin.** A large proportion of high school failures in Latin almost certainly could be avoided if an adequate system of guidance were employed to direct students out of the field when the essential aptitudes are missing. One of the most important functions of tests in the high school is to yield measures which are prognostic of a student's probable success in the various subjects. Unfortunately not much productive work has been done in Latin

prognosis. The high pupil mortality in first-year Latin is well known to teachers. Obviously much more attention should be given to the problem of determining which students are capable of studying Latin with profit. A student likely to be successful in the study of Latin must possess certain aptitudes for the learning of languages as well as the ability to see abstract relations. Students possessing these latter abilities will obviously secure more substantial returns from their study of Latin than those who may be linguistically inclined, but who have no ability to broaden their outlook. Most prognostic tests in the foreign languages have not provided for this important element in success and are therefore unable to furnish a very reliable basis for predicting who should study Latin. In common with most other prognostic tests, they are much more successful in pointing out those who should *not* study the subject. The negative aspect of prognostic measurement is one of its most discouraging features.

The *Orleans-Solomon Latin Prognosis Test*, the only test described in this section, is designed to provide a means of determining before study is begun what success students may be expected to have in the study of Latin. It affords a very satisfactory basis for educational guidance in this field, and for the classification of students in the courses according to their ability. The test consists of nine subtests in six of which special lessons are presented covering the same content as appears in the subtest. The efficiency with which the skills in these study-test situations are learned is taken as an index to the student's specific aptitude for this subject. The first of the subtests covers a sampling of English derivatives of Latin words for which meanings are given. The eighth and ninth subtests are designed to check the student's ability to acquire vocabulary under reasonably typical learning conditions. The administration of the test requires practically one hour. The evidence on the validity of the prognosis secured by the use of this test indicates that it is as valid as most similar types of instruments in related fields.

240. **Remedial Instruction in Latin.** The major areas of remedial instruction in the subject of Latin are those of oral and silent reading, vocabulary and grammar. Before effec-

tive remedial instruction in Latin is possible, a careful analysis and diagnosis of difficulties must be made in at least these three areas. Latin reading tests of the work-study type will do much to reveal reading deficiencies. Such reading tests may be of much the same character as those for silent reading in English. Such tests should be analytical within the limits of the subject, measuring such qualities as reading comprehension at the word level, comprehension in sentences and in larger units, reading rate, use of Latin books, and methods of study. There is also a need for oral reading tests in Latin which will function much like the Gray Oral Reading Tests in English.

Vocabulary tests such as the *Pressey Technical Vocabularies of Public School Subjects* which are described in Section 247 of this chapter are of considerable value in determining whether the student is deficient in his working vocabulary and basic Latin concepts. Syntax frequently constitutes a difficulty for the student in Latin. Because much of the syntax of Latin is not unlike that found in English grammar, drill on principles of English grammar should constitute an excellent basis for the work on Latin syntax.

Teachers will find many suggestions for the preparation of remedial exercises in the Latin tests which have been described in this chapter. The following classification of tests may be helpful in identifying the specific skills for which the test exercises are best adapted.

For Reading (*Comprehension*)
- Henmon Latin Tests (Series in Comprehension)
- Powers Diagnostic Latin Test
- Thorndike-McCall Reading Test
- Ullman-Kirby Latin Comprehension Test

For Composition
- Composition Scales
- Godsey Latin Composition Test

For Vocabulary
- Henmon Latin Tests (Series on Vocabulary)
- Powers Diagnostic Latin Tests
- Pressey Technical Vocabulary Tests (Latin)
- Stevenson Latin Vocabulary Test
- Stevenson-Coxe Latin Derivative Test
- Section of New York Latin Achievement Test

For Grammar-Syntax

Hutchinson Latin Grammar Scales
Kirby Grammar Test
Lohr-Latshaw Latin Form Test
Powers' Diagnostic Latin Test
Pressey Test in Latin Syntax
Parts of New York Latin Achievement Test

Numerous work-books are also available for assisting the teacher in his remedial work. While some of these may not be usable in their entirety, many are suggestive of the type of remedial material the teacher may prepare.

III. Measurement in the Modern Languages[7]

Tests in the modern foreign language fields probably do not differ significantly in the types of services rendered from tests in numerous other subjects. However, the following values suggested by Henmon[8] make the functions of foreign language tests so specific that they are reproduced here:

1. Tests set standards of accomplishment at different levels of training in objective, realistic, and comparable terms.
2. Tests make possible more accurate comparisons of attainment in different schools and classes under different methods and conditions.
3. Tests serve as a means of classification and placement to secure homogeneity in classes in terms of actual achievement in the languages rather than in terms of time spent in study.
4. Tests furnish instruments of analysis for investigation of the effects of varying ages, intelligence levels, methods, curricula, and objectives.
5. Tests aid in diagnosing deficiencies and locating them for definite remedial exercises.
6. Tests aid in defining in more specific terms the immediate objectives of instruction.

241. Measurable Elements in the Modern Languages. The four immediate objectives of instruction in the foreign languages, namely, the development of the ability to read, to write, to speak, and to comprehend the language, are of significance to the test-maker in this field. They obviously

[7] All tests mentioned in this chapter are listed with publishers in the *Appendix*, pages 585-595.
[8] Henmon, V. A. C., *Achievement Tests in the Modern Languages.* Vol. V, Modern Language Study. New York: The Macmillan Company, 1929.

determine the character of the tests to be used in the classroom. If these general abilities are to be measured effectively, the tests must be broken into many parts, each capable of measuring specific elements. In the *Modern Language Study* five of these abilities required in the acquisition of a foreign language were chosen as the basis for a test-construction program. Tests in vocabulary, grammar, silent reading, and written composition in the several modern language fields were actually constructed. Tests in aural comprehension, designed to be prepared on phonograph records in order to insure uniformity in administration, were projected but were not developed as a part of the *Modern Language Study*. Since that time tests of this type have been prepared on a commercial basis. The following suggested list of measurable elements includes the five emphasized in the *Modern Language Study*:

1. Vocabulary
2. Grammar
3. Silent reading
4. Written comprehension
5. Aural comprehension
6. Translation into English
7. Translation into the foreign language
8. Oral composition test
9. Oral reading test
10. Pronunciation test
11. Geographical and historical perspective, general cultural background in the homeland of the language
12. Common phrases, quotations from the language used in English writing and speech

Unfortunately, tests suitable for the measurement of all of these skills have not been devised as yet.

242. **Achievement in French.** Measurement of achievement in French employs techniques quite similar to those used in English and in the other modern languages. The emphasis is definitely upon testing the student's familiarity with the reading and speaking vocabularies of the language, the pronunciation and enunciation of words and idioms, the spelling of words, and certain word rules of construction with their exceptions. A few of the more commonly used

tests for the measurement of achievement in French are described briefly in this section.

The *American Council Alpha French Test* is a comprehensive test for use at the high school and college level. It was prepared originally for use in the *Modern Language Study.* While the two parts of the test together give a very complete measure, either part may be used alone. *Part I* consists of a vocabulary test of fifty words, and a grammar test of fifty items. *Part II* comprises a silent reading test and a composition test. The grammar test items, arranged in ascending order of difficulty, are those most common to widely used texts. The completion type of exercise is used. The comprehension test is constructed along the same general plan as was followed in the *Thorndike-McCall Silent Reading Scales.* The paragraphs are arranged in order of increasing difficulty, the student being required to write his answers in English. The composition test requires the student to write a composition in French, on a subject suggested by a picture which is included on the last page of the test.

This test measures with a high degree of reliability the four abilities in French as described in this chapter, and is undoubtedly among the best present available measures of achievement in French. Tables of percentile norms for both high school and college students are provided. Two equal forms of the test are available. The working time for each part is forty minutes.

The *American Council Beta French Test* is similar to the Alpha test with the exception that it is on a somewhat lower level and is to be preferred for the testing of first year students of French. This elementary test was devised for use in the *Modern Language Study.* The test consists of three parts, vocabulary, comprehension, and grammar. It is designed to provide junior and senior high schools with more reliable, valid and comparable measures of achievement in French than are afforded by the usual French examinations. Norms, based on over 10,000 students, are given. The test requires ninety minutes of working time.

The *American Council Grammar Test* was devised for use in the *Modern Foreign Language Study.* It was prepared as an alternate or substitute for the grammar part of the

American Council Alpha and Beta French tests, or other tests, for those persons who prefer the multiple choice technique. The test consists of fifty English sentences, each followed by four alternative translations, only one of which is correct. Forms A and B are not only equivalent but are supplementary, so that the two taken together constitute a very comprehensive examination. The time for administering is twenty-two minutes.

The *Broom-Brown Silent Reading Test in French* is designed to determine a student's ability to read French silently and to understand what he has read. The test is made up of a series of 20 paragraphs ranging from easy to difficult and covers a variety of reading materials. It is practically self-administering, and is convenient for testing large numbers. It requires twelve minutes of working time. The score is the number of correct answers based upon the paragraphs read.

The *Columbia Research Bureau French Test* is a reliable and valid achievement test designed for use in high-school and college classes. Two equal forms of the test are available, each covering vocabulary, comprehension, and grammar. The vocabulary and other elements in the test have been carefully selected to include material almost certain to be found in typical courses in the subject. Norms for each part of the test are based upon returns from more than twenty thousand high school students. The working time for the test is ninety minutes.

The *Handschin Modern Language Tests, French* comprise two different tests, Test B for silent reading, and Test A for comprehension and grammar. The silent reading tests may be used with students in the first or second year of French, while the comprehension and grammar tests are intended for students in their first year of French. Five minutes are allowed for the silent reading tests and ten minutes for the comprehension and grammar tests. The silent reading test is made up of a story of 192 words in French. The student is asked to read this story as rapidly as possible being sure to get the meaning. He is told that he will be asked to answer questions, and is given an opportunity to see the questions before the paragraph is read, but is not permitted to consult

the questions during the time of reading, or the paragraph when he is writing the answers to the questions. The scoring provides a measure for rate of reading as well as for comprehension.

Test A contains six easy French sentences which are to be studied for five minutes for the purpose of reproducing them correctly in French. The sentences have certain words omitted. The student is expected to supply the missing words in French. After these sentences have been reproduced, the student is directed to rewrite them "in the third person, plural, past indefinite tense," changing all words to the proper form. These tests possess some diagnostic value. In addition to providing a measure of comprehension, the tests will aid in determining a student's knowledge of the different forms of nouns, verbs, articles, adjectives and particles.

Henmon's French Tests originally comprised four tests, Tests 1 to 4, but Test 1 is now out of print. Each test contains a vocabulary test of 50 words and a sentence test of ten sentences. All the words used in the tests are common to twelve widely used beginners' textbooks. As the three different tests are of the same grade of difficulty, any one can be used first. They can be given at any time to students in any year of high-school French and will serve as a reasonably valid basis for grading attainment. The working time for the test is twenty minutes.

The *Sammartino-Krause Standard French Tests* cover vocabulary, grammar, and comprehension in each of the two parts. The vocabulary section is composed of words picked from nine standard textbooks in French. These words were then checked with Henmon's *French Word Book* (which gives the frequency of words used in 400,000 French running words), with Ward's *Minimum French Vocabulary Test Book*, and with Wood's *Comparative Study of the Vocabularies of Sixteen French Textbooks*. The grammatical items were derived from nine French textbooks, each item being common to all of these books. Probably any textbook used includes all of the items selected for the test. Several courses of study were also used in checking the items of the test. The comprehension part of the test probably

functions in a valid manner. English is used in questioning the student about his understanding of the French selections he has read. His answers are recorded in English. The entire test is easily administered and scored. Norms are provided for vocabulary, grammar, and comprehension for each part for the first, second, third, and fourth semesters of French. The working time for Part I is twenty-eight minutes; Part II, thirty-two minutes.

The *Twiggs French Vocabulary Test* is available in a single form for all four years of the study of French. The test contains 50 isolated French words after each of which the pupil is to write the English meaning, and 100 French sentences in each of which one word is italicized. The pupil is required to write after each sentence the English meaning of the italicized word. Each word has assigned to it a credit value according to its difficulty. The time for the test is forty minutes.

243. **Achievement Tests in German.** Achievement in German is generally expressed in terms of the ability of the student to translate from the foreign language to English and the reverse, and his abilities in grammar and syntax. Progress which has been made in the measurement of achievement in German may fairly be credited to the work of test construction and evaluation carried through by the staff of the *Modern Language Study.* Three of the more popular tests in German are described here.

The *American Council Alpha German Test* gives a comprehensive view of achievement in German. It is similar in purpose, derivation, and scope to the *American Council Alpha French Test* which has been described above. For each part there are two equivalent forms, A and B. The working time for each part is forty minutes. Reliability for different parts for single years ranges from 82 to .98.

The *American Council on Education German Reading Test* comprises Scale A, Division 1, for the first and second years of German; Scale A, Division 2, for the second and third years of German. These scales were prepared for use in the *Modern Foreign Language Study.* The content was selected from standard works used during the first three years of secondary school German. Each scale is made up

of fifteen paragraphs written in German. Following each paragraph are statements, one or more of which are to be checked, which contain an idea that is in the paragraph or that can be derived from it provided the paragraph is comprehended correctly.

These scales measure the students' ability to read and understand the German paragraphs given in the scales. Each paragraph is evaluated. The working time is a class period.

The *Columbia Research Bureau German Test* in its scope and purpose is similar to the *Columbia Research Bureau French Test*, described above. It may be used in the first four years of high school or the first three years of college. There are two equivalent forms of the test, A and B. The working time is ninety minutes. The correlation coefficients (r_{AB}) for the two forms vary from .80 to .95 for the separate parts. For the total score the correlation is .96.

244. **Achievement Tests in Spanish.** Knowledge of vocabulary and grammar, and the ability to read and write the language are the chief elements of achievement in Spanish which are now capable of measurement. Five of the more commonly used tests in Spanish are described here briefly as examples of the different types.

The *American Council Alpha Spanish Test* is similar in scope, purpose, and derivation to the *American Council Alpha French Test* described above. It is designed for use in high school and college classes. For each part there are two equivalent forms, A and B. The working time for each part is forty-five minutes. Reliability coefficients for the different parts for single years range from .84 to .92.

The *American Council Beta Spanish Test* is elementary in character and is similar in scope, purpose, and derivation to the *American Council Beta French Test* described above. There are two equivalent forms, A and B. The working time is ninety minutes.

The *Columbia Research Bureau Spanish Test* is available in two forms, A and B. In scope and purpose it is similar to the *Columbia Research Bureau French Test* described above. Norms are available for each part separately as well as for the whole test. It may be used in the first four years

of high school or the first three years of college. The working time for the test is ninety minutes.

The *Contreras-Broom-Kaulpers Spanish Reading and Vocabulary Tests* are designed for use in high schools and colleges. The *Silent Reading Tests* are available in two forms, A and B; the *Vocabulary Tests*, in three forms, A, B, and C. The Silent Reading Tests are of the multiple-choice type. There are twenty paragraphs followed by questions with five possible answers, one of which is correct. All exercises are in Spanish. The Vocabulary Tests comprise 67 exercises for each form. In each exercise there are four Spanish words and four English words. The student is required to select and underline the Spanish and the English words which are most nearly similar in meaning.

The authors of these two series of tests claim they should be of value to teachers of Spanish in the following ways:

1. In classifying students into sections for the purpose of differentiating their rates of progress or for enriching the curriculum for the more able students.
2. In the determination of promotions and failures.
3. In the assignment of school marks, by furnishing a more objective and reliable basis than the teacher's unsupported judgments of achievement.
4. In setting up immediate educational objectives, i.e., checking the course of study to determine the amount and rate of instruction with given groups of students.
5. In comparing results secured by a teacher in different sections of her courses or by different teachers of similar courses in the same school or in different schools.

Norms for the Silent Reading Tests are provided separately for boys and girls having had one to eight semesters of Spanish and for college classes. Norms for the Vocabulary Tests are provided separately for boys and for girls having had one to six semesters of Spanish and for college classes. The working time for each test is about twenty-two minutes.

The reliability coefficient for the Silent Reading Tests is about .80 and for the Vocabulary Tests about .90.

The *Handschin Modern Language Test, Spanish* contains a silent reading test constructed along the same general plan as the *Handschin Silent Reading Test B, French* which

is described in an earlier section. This test is quite similar in scope and function to the Handschin tests in French.

245. Prognosis in the Modern Languages. Prognosis in the modern languages is as necessary and quite as practicable as it is in Latin. Many high school and college students fail in modern language study, some because they lack interest or intelligence, others because they lack the special abilities required for effective work in these fields. It is generally agreed that a large percentage of these failures could probably be avoided if an effective guidance program were operating to advise certain students away from the subjects for which they have no aptitude. A few tests capable of yielding the necessary type of prognostic information for giving accurate guidance in the study of the foreign languages are now available.

The *Foreign Language Aptitude Test* of the Iowa Placement Examinations undertakes to furnish a prognostic measure by sampling into four types of skills found to be highly related to success in the study of the foreign languages. A unique feature of this test is the use of *Esperanto* as the unfamiliar situation in which to test the student's ability to acquire new language skills. The test is similar in scope, purpose and derivation to the other aptitude tests of the Iowa Placement series described on page 405 in Chapter XVI and on page 467 of this chapter.

The Luria-Orleans Modern Language Prognosis test is designed to measure the ability of students to learn French, Spanish, or Italian. It is intended to be used before study of the language is begun. The results are useful as a basis for advising students whether to begin a foreign language, as a guide in classifying students, and as an aid in determining the effectiveness of instruction. The test contains a number of simple lessons with a test on each, covering certain fundamental principles and skills involved in learning a Romance Language. There is only one form of the test. The working time is seventy-six minutes.

The *Wilkins Prognosis Test in Modern Languages* is valuable in eliminating from modern languages courses those pupils who will probably be unsuccessful, and for classifying the students who can learn the language on the basis of

their apparent abilities. This series contains six different tests as follows:

Test 1. *Visual motor* (seeing and writing)
A French or Spanish sentence written upon a flash card is shown to the student for five seconds.
The student is then required to write the sentence.

Test 2. *Aural motor* (hearing and writing)
A French or Spanish sentence is read to the student. He is then given 10 minutes to write it as he heard it.

Test 3. *Memory*
The student is given two minutes to study ten words in French or Spanish with their English equivalents, after which he is required to write them.

Test 4. *Grammar Concepts*
The student is given 15 minutes to change the forms in a group of English sentences.

Test 5. *Visual-oral* (seeing and speaking)
The student is given three seconds in which to observe an English sentence on a flash card and then to report it to the examiner.

Test 6. *Aural-oral* (hearing and speaking)
The student is required to repeat a French or Spanish sentence which the examiner pronounces to him.

The first four tests are group tests, while tests five and six must be administered individually. The pupil's composite score on the six tests is taken as an index of his ability to learn a modern language. To be truly prognostic, these tests should be given before the student has had an opportunity to acquire a knowledge of the foreign language. The tests may be used also for the classification of students into groups for purposes of modifying instruction in the language according to the ability of the group.

246. **Remedial Work in Modern Languages.** The problems of diagnosis and remedy in the modern foreign languages are essentially the same as those in Latin. Reading, pronunciation (accent), vocabulary (idiom), and grammar constitute the major remedial areas. Naturally a careful diagnosis of possible causes of difficulty must be made. The results of intelligence tests will be helpful in locating general causes of trouble. Prognostic and aptitude tests will also be useful in identifying students who apparently lack the special abilities upon which successful work in the foreign language

seems to depend. Tests with diagnostic or analytical power or special tests for relatively narrow skills must also be used. Such tests as the *Sammartino-Krause French Tests* for determining control of grammar, vocabulary and comprehension, the *Pressey Technical Vocabularies* for determining vocabulary deficiencies, the *American Council Tests* for analysing abilities in vocabulary, grammar, silent reading and composition should assist the teacher in discovering those elements of foreign language study which require additional emphasis.

Informal objective tests in vocabulary, grammar, composition, and reading comprehension will also be useful. The teacher will find a large amount of material and many suggestions on new-type tests in the modern languages in Volume I of the *Modern Language Study.*[9] The *Modern Language Study* also summarizes much material on teaching methods, vocabulary and idiom lists, as well as results of special studies which should prove invaluable to the teacher of the foreign languages.

The analysis of causes of pupil failure is indispensable to remedial work. The results of such an analysis may readily provide the basis for preventive work designed to reduce the number of failures. Low mentality, large student turnover, bad health conditions, poor study conditions, ineffective methods of study, lack of co-operation between teacher and student, and classes which are too large or too heterogeneous are certainly some of the factors which necessitate an effective remedial program. Without doubt one very potent reason for student failure in the foreign languages is the lack of suitable instruction in effective methods of study. Another important factor may be the failure of teachers to understand the varying interests, needs, and future intentions of their students, and to adjust their courses accordingly. The accompanying list of study-helps is offered merely as a suggestion which the wide-awake teacher will readily elaborate.

1. Vocabulary study —
   - *a.* Review of former vocabularies
   - *b.* Correlation of new words
   - *c.* Learning of idioms

[9] Wood, Ben D., *New York Experiments with New-Type Modern Language Tests.* New York : The Macmillan Company, 1927.

   d. Practice in use of a, b, c in original sentences
   e. Applying knowledges gained in original ways
2. Grammar study —
   a. Practice in analysis and discrimination
   b. Discovery of short cuts
   c. Independent thinking
3. Reading —
   a. Daily review
   b. Drill in rapid reading for quick comprehension
   c. Drill in association of ideas
   d. Reading for the idea
   e. Reading aloud selected portions
   f. Checking of reading results
   g. Exercise in reading foreign magazines and newspapers
4. Composition —
   a. Concentration on such essentials as position, agreement, gender, mood, and accentuation
   b. Checking of results by studying corrected work
   c. Emphasis upon accuracy

247. **Miscellaneous Language Tests.** The three types of foreign language testing material described in this section do not seem to fit into any one of the foregoing sections.

The *Iowa Every-Pupil Test in Latin* is one of the series of thirteen subject-matter tests used in the Iowa Academic contest each year. The content and the structure of these tests changes from year to year so that it is impossible to describe them adequately. For the 1935 series the Latin tests are made up of 179 items arranged in four subtests. The testing time required is sixty minutes. This complete series of tests is described in more detail in Chapter XXII.

The *Iowa Placement Examinations* offer three types of tests for use in the foreign languages at the college level. The *Foreign Language Aptitude* test is described in Section 245. The *French Training Test* comprises four parts as follows:

Part 1.	French to English Vocabulary
Part 2.	French Grammar
Part 3.	French Idioms and Tenses of Verbs
Part 4.	French Reading Comprehension

The *Spanish Training Test* is also arranged in four parts paralleling in general structure the *French Training Test.*

The function of these tests is to determine the extent to which previous instruction in the subject has proved to be effective. They are used for placement and sectioning purposes in college classes in which students have already had previous instruction.

The *Pressey Technical Vocabularies of the Public School Subjects* are available in fifteen different subjects. The vocabulary lists in French, German and Spanish offer real possibilities for testing the basic concepts in these subjects. It is suggested that the student go through the lists and systematically master each word. The author states that "these vocabulary lists are more than mere lists of words; they are catalogues of important concepts of the subject. So the lists may be used as diagnostic tests in determining the specific weaknesses of a class or an individual pupil in the subject. That is, if the teacher will find out the terms which a pupil does not know, in the vocabulary . . . she will have a very helpful indication as to the topics on which that pupil needs help. The list can be of great service in thus diagnosing weaknesses, and guiding the teacher in her remedial instruction."[10]

Discussion Exercises

1. What distinguishes an immediate objective of instruction from an ultimate objective?
2. What important changes have been made recently in the objectives of modern language teaching?
3. List for Latin, French, German, and Spanish the most important immediate outcomes of each. Are these outcomes significantly different for any one language? What does this suggest to you?
4. Make a list of the measurable qualities which seem to be common to all foreign languages.
5. Point out the specific differences in the types of measurement techniques which are employed in the foreign languages and in English.
6. List the pro and con arguments for the continuation of instruction in the modern foreign languages in the high school.

[10] Pressey, L. C., *Technical Vocabularies of the Public School Subjects.* Bloomington, Illinois: Public School Publishing Company, 1923.

7. Examine the available tests in the fields of Latin, French, German, and Spanish and select one test suitable for high school use in each field.
8. In your judgment what are the greatest weaknesses of foreign language tests?
9. Outline your idea of what specific skills an analytical test in one of the modern foreign languages should cover.
10. Select some one aspect of achievement in a modern language and outline at least three suggestions for remedial or corrective lessons.

Selected References

Alexander, H. B., "Modern Languages and Western Civilization," *Modern Language Forum*, Vol. 14, No. 1, January, 1929.

Blackhurst, J. H., "Aims of Modern Language Instruction," *Educational Administration and Supervision*, Vol. 12:625-629, December, 1926.

Brueckner, L. J., "The Status of Certain Basic Latin Skills," *Journal of Educational Research*, Vol. 9 : 390-402, May, 1924.

Buchanan, M. A., and MacPhee, E. D., *An Annotated Bibliography of Modern Language Methodology*. Publications of the American and Canadian Committees on Modern Languages, Vol. II. New York : The Macmillan Company, 1927.

Cerf, Barry, "Aims in the Teaching of Modern Languages," *Modern Language Journal*, Vol. 6 : 419-440, May, 1922.

Classical League, The American, *The Classical Investigation*. Princeton, N. J. : Princeton University Press, 1924.

Cole, Robert D., *Modern Foreign Languages and Their Teaching*. New York : D. Appleton and Company, 1931.

Coleman, Algernon, *The Teaching of Modern Foreign Languages*. Publications of the American and Canadian Committees on Modern Languages. Vol. XII. New York : The Macmillan Company, 1929.

Eddy, Helen M., *Instruction in Foreign Language*. Monograph 24, National Survey of Secondary Education, Bulletin No. 17. Washington, D. C. : United States Department of Interior, Office of Education, 1932.

Game, J. B., *Teaching High School Latin*. Chicago : University of Chicago Press, 1925.

Henmon, V. A. C., *A French Word Book Based on 400,000 Running Words*. Bureau of Educational Research Bulletin, No. 3, University of Wisconsin, 1924.

Henmon, V. A. C., *Achievement Tests in the Modern Foreign Languages.* Publications of the American and Canadian Committees on Modern Languages, Vol. V. New York: The Macmillan Company, 1929.

Jordan, J. N., "Prognosis in Foreign Language in Secondary Schools," *School Review,* Vol. 33 : 541-46, September, 1925.

Kelsey, F. W., *Latin and Greek in Secondary Education.* New York: The Macmillan Company, 1927.

Krause, C. A., *The Direct Method in Modern Languages.* New York : Charles Scribner's Sons, 1916.

Morgan, B., "The Place of Modern Foreign Languages in the American High School," *School and Society,* Vol. 27 : 185-192, February 18, 1928.

O'Shea, M. V., *The Reading of Modern Foreign Languages.* United States Bureau of Education, Bulletin, No. 16, 1927.

Patterson, A. S., "A Correlation of Aims and Methods in Foreign Language Teaching," *Modern Language Journal,* Vol. 7 : 385-402, April, 1923.

Thomas, Calvin, "Good and Bad Reasons for Studying Modern Languages in School," *Modern Language Journal,* Vol. 5 : 1-11, October, 1920.

Wood, Ben D., *New York Experiments with New-Type Modern Language Tests.* New York: The Macmillan Company, 1927.

CHAPTER XIX

MEASUREMENT IN THE FINE ARTS

THE purpose of this chapter is the discussion of the following possibilities of measurement in the fine arts:

- *a*. Social and educational significance of the arts.
- *b*. Educational emphasis on art and music.
- *c*. Basic elements of musical talent.
- *d*. Measurement of musical accomplishment.
- *e*. Measurement of artistic talent.
- *f*. Measurement of artistic ability.

The objective measurement of achievement in the fine arts is a relatively recent accomplishment. In fact, it is so recent that there is still an echo of protest from certain conservative artists that artistic production does not lend itself to measurement. In spite of this feeling, however, recent years have seen much progress in these fields. This is proper, for certainly in these cultural subjects is to be found much of the best that our educational program affords. With the pendulum swinging away from the industrial-vocational emphasis on life's activities in the direction of greater individual leisure for the cultural pursuits, the need for a better understanding of the content, aims, and methodology of these artistic subjects is greater than ever before.

There is perhaps a certain advantage in the fact that developments in the measurement of the fine arts have taken place slowly and quite recently. In general, research techniques have improved, with the net result that the problems of measurement in these fields have been more critically analyzed and attacked with more refined instruments. The very critical and carefully controlled research of such workers as Seashore, Schoen, Stanton, Kwalwasser, and Dykema in the psychology and pedagogy of music, and the

work of Thorndike, Ayer, Meier, Manuel, McAdory, and Whitford in the field of art are evidences of the influence of this point of view.

I. Measurable Qualities in Music

Measurement in music takes two major lines of approach. The first is the determination of basic aptitudes. Here as in other subjects, the techniques and instruments used are psychological. Such instruments have been previously described in this volume as tests of specialized intelligences, since they have to do with the determination of inherited tendencies to respond in certain manners to specific types of training stimuli. Accomplishment in music depends to such a large degree upon the existence of special aptitudes that this phase of measurement must be emphasized. The mere existence of aptitude in music is in no sense an index to musical accomplishment, however. The second approach to the problem is pedagogical and is based upon the use of achievement tests for the three-fold purpose of measuring the knowledges, skills, and appreciative aspects acquired as a result of training. As Kwalwasser[1] points out, "regardless of the talent possessed, one must have the will to succeed or little is attained. . . There are a vast number of reasons why an individual of superior endowment may realize but a very small return on his native musicianship."

248. **Aims and Outcomes in Musical Education.** The 1921 report of the *Educational Council of the Music Supervisors National Conference* presented a course of study in music setting forth the aims and outcomes of music education in a form which has not been significantly improved upon since that time. While the aims and attainments are repeated for each grade in the report, only the general aims appropriate for the end of the eighth grade are summarized here. All aims and outcomes of prior grades are incorporated in this summary. Items *a* to *k* inclusive represent aims to be attained prior to the seventh grade. Items *l* to *s* are specific objectives of the seventh and eighth grades.

[1] Kwalwasser, J., *Tests and Measurements in Music.* Boston : C. C. Birchard and Company, 1927.

Aims (At End of the Eighth Year)

a. To continue the development of free and beautiful singing of songs
b. To acquire an increasingly wide musical experience
c. To develop increasing power of eye and ear in correlation
d. To develop power to listen for musical beauty as well as for musical knowledge
e. To develop increased power to sing at sight
f. To establish two-part singing
g. To develop increasing practical knowledge of the tones of the chromatic scale and power to use them
h. Extension of knowledge of the tonal and rhythmic material of music appropriate to the grade
i. To develop a fair degree of power to sing unison songs at sight with words, and an elementary degree of power to sing two-part songs at sight with words
j. To begin the development of three-part, treble-voice singing
k. To develop ability to deal practically with the minor mode
l. To develop concerted singing in the direction of mass practice as well as to continue the usual classroom sight-singing and part-singing
m. To recognize the birth of new affective (emotional) states in the pupils, due to their awakening sense of the relationships of human life and the emotional aspects of these relationships ; and to utilize the best of these qualities of feeling as agencies toward the reinforcement and upbuilding of fine and strong elements of character
n. To articulate more closely for the pupils, individually and collectively, the musical interests and activities of the school with those of their homes and their community
o. To recognize and encourage the special interest that pupils of this age have in the mechanism, technique and use of musical instruments
p. To recognize and encourage individual musical capabilities, as a feature of an avocational as well as a vocational stage of development
q. To pay special attention to the diverging needs of the voices of pupils
r. To strengthen and extend technical knowledge and capability with reference to tonal and rhythmic elements and features of staff-notation and sight-singing
s. To add to the appreciation of the formal elements in music an appreciation of the moods characteristic of romantic and modern music

Attainments

a. Ability to sing well, with enjoyment, at least 30 unison, two-part, and three-part songs, some of which shall be memorized
b. Ability of 90 per cent of the pupils to sing individually, freely, correctly, and without harmful vocal habits not less than ten of the songs sung by the class as a whole

c. Ability to sing at sight, using words, a unison song of hymn-tune grade; or using syllables, a two-part song of hymn-tune grade, and the easiest three-part songs; these to be in any key; to include any of the measures and rhythms in ordinary use; to contain any accidental signs and tones easily introduced; and in general to be of the grade of folk songs such as "The Minstrel Boy." Also knowledge of the major and minor keys and their signatures

d. Ability of at least 30 per cent of the pupils to sing individually at sight music sung bv the class as a whole

e. Ability to appreciate the charm of design in songs sung; to give an account of the salient features of structure in a standard composition, after a few hearings of it; to identify at least the three-part song form from hearing; to recognize and give titles and composers of not less than 20 standard compositions studied during the year

f. Ability to sing well, with enjoyment, a repertory of 25 to 35 songs of musical, literary, community, national or other worthy interest

g. Ability to sing at sight part songs of the grade of a very simple hymn

h. Knowledge of all essential facts of elementary theory sufficient to enable 75 per cent of the students to give a correct explanation of any notational features contained in the pieces of average difficulty in the standard books of music for the eighth year

i. Further progress in recognition of the relations, agreements, dependencies of tones and tonal groups, that give music its strength and interest; pleasure in good music.[2]

There is little doubt that the foregoing statement of objectives and attainments in music has in the period since its appearance exerted a significant influence on the place of music in the schools. The current interest in music festivals and contests in which school-trained musicians appear is tangible evidence of this influence. Prominent among the features of the modern school program is the rich course of study in music which is offered to high school pupils to equip them for avocation or vocation. As a result of greatly improved equipment and instruction made possible through community interest and co-operation, many of these courses in music are exerting an influence reaching far beyond the confines of the school. In one midwestern state alone over 5,500 high school musicians recently took part in the annual music festival and contest. Such interest in music education

[2] Items *f* to *i* represent attainments specifically set for the seventh and eighth grades.

is certain to make itself felt in the culture of a state or a nation.

II. Measurement of Musical Abilities and Achievement

249. **Basic Musical Talent.** Tests of musical aptitude are designed to measure those innate musical capacities which constitute the individual's musical inheritance. Aside from the sheer physical endowment which certain types of musical expression demand, there are certain more or less psychological factors which determine an individual's musical talent. The identification of these factors calls for an unusually critical analysis. Without doubt one of the most extensive research programs ever undertaken for the purpose of isolating the elements of native capacity in a special field is the development by Seashore[3] and his students over a period of twenty-five years of his *Tests of Musical Talent.* According to the author's own statement quoted from the manual for these tests, "these measures of musical talent comply with the following conditions: they are based on a thorough analysis of musical talent; they are standardized for content that does not need to be changed; they give quantitative results which may be verified to a high degree of certainty; they are simple and as nearly self-operating as possible; they are adapted for group measurements; they take into account practice, training, age, intelligence; they have a two-fold value in the concrete information furnished, and in the training and pleasure gained from the critical hearing of musical elements."

These tests, six in number, are designed for use not earlier than the fifth grade. Experience indicates that this is practically the earliest age at which such group measures may be taken, and that it is a suitable age at which to begin making serious arrangements for a musical education if it appears desirable. The six elemental abilities measured by the tests in their present form are (1) Sense of Pitch, (2) Sense of Intensity, (3) Sense of Time, (4) Sense of Consonance, (5)

[3] Seashore, C. E., *Measures of Musical Talent, Manual of Instructions and Interpretations.* Chicago: C. H. Stoelting Company.

Tonal Memory, and (6) Sense of Rhythm. The test stimulus to which the student responds is supplied by means of double-faced disc records for phonographic reproduction. The tests may be administered to groups, the size of the group depending somewhat on the acoustical qualities of the room. Naturally the stimulus must be heard clearly at all times. Further details concerning the administration and interpretation of these tests may be obtained by reading Seashore's *Psychology of Musical Talent.*[4] These tests are soon to be made available in a completely revised form in which advantage is taken of the many constructive criticisms which have resulted from the numerous research studies in which they have been used. In this revised form a new measure of the Sense of Timbre will be added.

Norms are provided for grades five and eight and for adults. These tests cover only a limited range of the whole field of musical ability, but without question they are valuable for the identification of students who would profit from a musical education only indifferently because of deficiencies in one or more of the six capacities measured.

The *Kwalwasser-Dykema Tests*[5] are quite similar in form and function to the *Seashore Musical Talent Tests.* There are ten tests in this series, requiring five phonographic records. The tests are designed for use in grades four to twelve. The elements shown in Table 59 are measured by means of alternate response testing techniques utilizing the indicated numbers of items.

The student who is particularly interested in the measurement of basic musical talent will naturally wish to secure a critical and comparative evaluation of the *Seashore Tests of Musical Talent* and the *Kwalwasser-Dykema Tests* since these two represent the most useful measures of musical talent available at the present time. For this purpose a study by Paul R. Farnsworth reported in the May, 1931, *Genetic Psychology Monograph* is very illuminating.

Quite in contrast with the Seashore and Kwalwasser-Dykema tests is a new paper-pencil test of musical talent,

[4] Seashore, C. E., *The Psychology of Musical Talent.* New York: Silver, Burdett and Company, 1919.

[5] *Kwalwasser-Dykema Tests.* Published by Carl Fischer, Inc., Chicago.

TABLE 59

CONTENT OF KWALWASSER-DYKEMA MUSIC TESTS

Test No.	Test Name	Items in Test
1	Tonal Memory	25 pairs
2	Quality Discrimination	30 pairs
3	Intensity Discrimination	30 pairs
4	Tonal Memory	30 tonal patterns
5	Tone Discrimination	25 3-note units
6	Rhythm Discrimination	25 pairs
7	Pitch Discrimination	40 items
8	Melodic Taste	10 pairs
9	Pitch Imagery	25 tonal patterns
10	Rhythm Imagery	25 rhythm patterns

the *Drake Musical Memory Test.*[6] The author states that the "discovery of innate talent is the chief aim of this test. It does not measure acquired information or progress in music as achievement tests attempt to do, but measures *capacity* for musical achievement the same as the so-called intelligence tests measure capacity for school achievement." In its essentials, it is another test of musical memory. It remains to be seen whether or not it is better than other similar tests of this type. The test itself is designed for use with any group, of any age, with or without musical training. It can be given to eight-year-old children. A twenty-five minute testing period is required for each form taken singly. If both Forms A and B of the musical scores for the melodies are used for the sake of additional reliability, a forty minute period is required. The score sheet, a copy of which is reproduced here, gives detailed directions for the administration of the test.

250. **Musical Achievement.** The statement of the aims and actual attainments presented in the preceding section affords a reasonably definite basis for the general validation of tests of musical achievement. Among the tests of this type which sample rather widely into the specified instructional outcomes is the *Kwalwasser-Ruch Test of Musical*

[6] Drake, Raleigh M., *Musical Memory Test.* Bloomington, Illinois: Public School Publishing Company, 1934.

Accomplishment.[7] The battery itself consists of ten tests measuring the specific aspects of musical accomplishment shown in Table 60.

Table 60

CONTENT OF KWALWASSER-RUCH TEST OF MUSICAL ACCOMPLISHMENT

Test	Number of Items	Type of Item
1 Knowledge of Musical Symbols and Terms	25	Multiple-Choice
2 Recognition of Syllable Names	25	Identification
3 Detection of Pitch Errors in a Familiar Melody	5	Identification
4 Detection of Time Errors in a Familiar Melody	5	Identification
5 Recognition of Pitch Names	20	Identification
6 Knowledge of Time Signatures	10	Multiple-Choice
7 Knowledge of Key Signatures	15	Identification
8 Knowledge of Note Values	5	Multiple-Choice
9 Knowledge of Rest Values	5	Identification
10 Recognition of Familiar Melodies from Notation	10	Identification

The reliability of measurement of the test is reported as .97 based on 167 cases. Percentile norms based upon 5,414 cases are provided for grades 4, 5, 6, 7, 8 separately, and 9 to 12 combined. The ten tests require a total of 40 minutes testing time.

Another aspect of achievement in music which is important enough to deserve more attention is the measurement of music information and appreciation. The *Kwalwasser Test of Music Information and Appreciation* is an attempt to measure the extent to which these important outcomes of the course of study in music are attained. This test, which is standardized for use with high school and college students, is composed of nine objective sub-tests. The first five of

[7] Kwalwasser, J., and Ruch, G. M., *Kwalwasser-Ruch Test of Musical Accomplishment.* Iowa City : Bureau of Educational Research and Service, Extension Division, University of Iowa, 1927.

the sub-tests measure the student's information on (1) the classification of artists according to the field of music in which they are most prominent, (2) the nationality of composers, (3) the composers of famous compositions, (4) the classification of composers according to their works, as songs, piano, orchestra or opera, and (5) the general knowledge of composers and compositions. Knowledge of instrumentation is tested in three ways; namely, (6) the production of tones on orchestral instruments, (7) classification of orchestral instruments, and (8) general knowledge of instrumentation. Test 9 is designed to measure the student's general knowledge of music structure and form. Percentile norms are available.

The *Beach Standardized Music Tests* were among the real pioneers in the measurement of musical achievement. Many of the elements measured by these tests are recognized among the tests of more recent development. The following qualities are scheduled for measurement by the test:

1. Knowledge of essential fact of musical notation
2. Ability to hear and distinguish the component parts of music, namely the elements of time and tune both in isolated form and in melodies
3. Aural recognition of the structural elements of music fundamentally necessary for intelligent appreciation
4. Pitch discrimination
5. Musical memory
6. Sight-singing through indirect methods
7. The writing of music

The administration of the test is much more difficult than is indicated by the qualities it attempts to measure. It requires a forty minute testing period. The administrator of the test must play and sing many of the ear-training exercises, a difficulty which introduces many serious variations in the results. Kwalwasser, in his discussion of *Tests and Measurements in Music*, criticizes this test also for its difficulty in grading and for the fact that no norms are available.

Another test which has had considerable use is the *Torgerson-Fahnestock Music Test.* Modern music supervisors criticize it mainly because of the unduly large amount of emphasis given to the writing of music, a practice which

many supervisors consider an inexcusable waste of the pupil's time.

The *Providence Inventory Test in Music*,[8] one of the more recent developments in the measurement of music accomplishment, is standardized for use in Grades 4 to 9 inclusive. The authors state that the test represents items usually taught in the elementary course of study in music, but they do not give any definite basis for the selection of the items. Age and grade norms are furnished, as are also suggested distributions to be used for assigning letter ratings to pupils on the basis of test scores. This would appear to be a somewhat doubtful proceeding in view of the uncertainty of the curricular content emphasized in a particular classroom. Table 61, showing the number and names of the test parts, the number of items, the testing time, and the form of the exercises used in each of the ten tests gives a fair idea of the subject matter covered by the tests.

TABLE 61

CONTENT OF PROVIDENCE INVENTORY TEST IN MUSIC

Test No.	Test Name	No. Items	Testing Time	Form of Exercise
1	Naming Notes	12	2	Recall
2	Placing "Do"	9	2	Recognition
3	Naming Note Values	4	1	Multiple-Choice
4	Naming Key Signatures	11	2	Recognition
5	Naming Measure Signatures	6	2	Recall
6	Naming Rest Values	4	1	Multiple-Choice
7	Naming Syllables	12	6	Recognition
8	Naming Melodies	7	5	Recognition
9	Naming Syllables (Bass clef)	9	5	Recognition
10	Naming Symbols	9	2	Recall

Numerous other tests of musical accomplishment are now available for general use. Doubtless certain of these may be of quality as excellent as any of those described in this

[8] Allen, Butterfield, and Tully, *Providence Inventory Test in Music.* Yonkers, New York: World Book Company, 1932.

chapter. No attempt is made to make this list inclusive. Only a few of the representative tests suitable for use in the junior or senior high school are discussed.

251. **Remedial Materials in Music.** Remedial instruction in music consists mainly in the application of further practice on the specific skills desired. Musical aptitudes, of course, do not lend themselves to corrective treatment. Certain of the phases of musical accomplishment respond to additional, or different treatment. Thus remedial teaching in this field becomes mainly a process of identifying certain habits and informations and applying to them the required further drill and instruction. As a matter of fact, there is available almost no material of definitely remedial nature in the field of music.

III. MEASUREMENT OF ART ABILITIES AND ACHIEVEMENT

252. **General Outcomes of Art Instruction.** Human expression in the form of art dates back to pre-historic man. It appeared in many forms such as painting, carving of wood and stone, modeling of pottery, etc. Historically, art has left a most important record. To quote from one of the most recent discussions of art education,[9] "To teach Art as historical information is to reveal the evolution and occasionally the degeneration or regression of the human race. When all else has vanished away we still have left for study the art which each country has produced." In addition to this informational outcome of art instruction these same authors suggest two other general objectives which appear to be important. These are appreciation and expression. It is quite probable that art appreciation is not necessarily taught, although real appreciation may be considered to rest to a large degree upon the broader aspects of information. There will still remain something in the truly artistic product which sheer information does not entirely explain. The third objective might be better expressed as exploration. Not many potentially great artists are discovered in the ele-

[9] Klar, Winslow, and Kirby. *Art Education in Principle and Practice.* Milton Bradley Company, 1933.

mentary-school or high-school classroom, but practically all the great artists there are have come up through this avenue. Not every one can express himself effectively in artistic form, but everyone has a right to explore for himself the fields of human expression in the hope that his own hidden talent may be uncovered. Art-talent and achievement tests have distinct contributions to make in this field.

253. **Specific Outcomes.** The elementary school art course faces the responsibility for bringing to the child a four-fold artistic experience. These categories of experience were suggested by Kirby[10] in a recent discussion of aims and tendencies in art education. The first is the *graphic* experience which expresses itself in representative drawing, illustrative and imaginative drawing, nature drawing, and other related forms. The second is the *thoughtful* experience involving the constructive, decorative phases of artistic expression. The third experience involves the *acquisition of motor skill* in expression. The fourth is the *emotional* experience which involves the appreciation of the arts.

A somewhat more specific expression of outcomes of instruction in art is given in the accompanying outline of specific outcomes adapted from a course of study covering the first six years of the elementary school. It will be noted that this course is organized around four groups of outcomes which are quite similar to the artistic experiences presented in the previous paragraph.

Outcomes of Art Instruction[11]

A. Fruitful Knowledge
- Functional information
- Practical relation of art to everyday life (clothing, home, town or city, etc.)
- Understanding of elements and principles of art and their adaptation to everyday use
- Knowledge of construction and industrial processes involved in art training
- Acquaintance with art of other countries

[10] Kirby, C. Valentine, "Aims and Tendencies," *Pennsylvania School Journal*, Vol. 77 : 501-2, April, 1929.

[11] Adapted from Outline for Course of Study in W. G. Whitford, *An Introduction to Art Education.* New York : D. Appleton and Company, 1929.

B. Attitudes, interests, and appreciations
Civic consciousness (civic pride)
Appreciation and understanding of beauty in modern products of all kinds
Interest in art museums, travel, and further study
Interest in the civic, domestic, and social service of art

C. Mental technique
Good taste, discriminating judgment, ability to select and choose wisely
Creative ability, originality, initiative, imagination, keen observation
Ability to analyze works of art and to understand the factors of beauty in production
Keener observation, beauty of nature and fine things of art

D. Right habits and skills
Constructive thinking and planning
Systematic organization
Practical technique
Co-ordination of mind, hand, and eye
Freedom and spontaneity
Order, neatness
Body and mind training
Self-activity
Worthy use of leisure time

The accomplishment of these specific outcomes of instruction in art in the elementary school provides an excellent basis for the general art course in the high school. The following statement of purposes of such a general art course is summarized and quoted from Chapter XII of Whitford's *Introduction to Art Education:*[12]

1. To develop appreciation (*visual*, *mental*, *enjoymental*) through
   a. The development of knowledge and understanding of art quality
   b. The development of a love for and an understanding of beauty in nature and in worthy works of man
2. To give exercise in creating beauty (*technique*, *motor*) through
   a. The development of ability to choose, and arrange fine objects for specific purposes
   b. The development of ability to originate, design and produce fine objects with respect to standards of excellence
   c. The increase in visual experience
   d. The development of powers of invention, imagination, observation

[12] *Ibid.*

254. Measurement of Artistic Talent. In addition to manual dexterity there are a number of attributes of artistic talent. Manuel[13] reports the experimental use of tests designed to measure the following special functions in an attempt to reveal ability in drawing:

1. Ability to represent by lines and areas, figures and objects observed
2. Ability to observe
3. Ability to select from a complex situation the most representative and most beautiful aspects
4. Memory for visual forms
5. Mental manipulation of visual forms
6. Co-ordination of hand movements with visual image
7. Ability to bring together into new artistic combinations the elements of different visual experiences
8. Ability to judge the beautiful in line, form, color, composition
9. Ability to discriminate differences in color
10. Ability to discriminate differences in visual magnitude
11. Acuity of vision
12. General intelligence

The reliable measurement of artistic talent depends upon the success with which these and other essential attributes underlying achievement in art can be identified and measured with valid instruments. It is significant that two of the more recent and better known measures of art talent have hit upon art judgment as a basic talent. An extensive program of research has been carried on by Dr. Meier in a definite attempt to isolate the importance of this judgment factor. His most recent results quite thoroughly substantiate his earlier findings with respect to the validity of the *Meier-Seashore Art Judgment Test*[14] as a measure of artistic talent.

In the *Meier-Seashore Art Judgment Test*, which may be given as an individual or as a group test, the student is confronted with 125 pairs of artistic specimens adapted from many sources. One of each pair of specimens has been changed in some specific element. The exact feature changed is specified in the record sheet on which the student records his reaction. A consideration of the complete series of paired specimens insures a comprehensive sampling of the

[13] Manuel, H. T., *Talent in Drawing.* Bloomington, Illinois: Public School Publishing Company, 1919.

[14] Meier, Norman C., *Studies in the Psychology of Art.* University of Iowa Studies in Psychology. Published as Vol. XLV, No. 1 of Psychological Monographs, 1933. Princeton, N. J.: Psychological Review Company.

various elements which appear to affect aesthetic judgments. According to the evidence obtained by the authors this test appears to measure the sensitivity of the individual to the effect which the composition as a whole produces on the observer. In order to give a more exact idea of the nature of the specimens and the accompanying record sheets, a single pair of the etchings is reproduced here (Figure 14) along with a brief sampling of five items from the Test Record Sheet. The pair of specimens reproduced here are used with item 125 in the record sheet. According to this item the position of the largest tree in the picture is the point for special consideration in judging this pair. The answer key indicates that the position of the tree is better in the left-hand specimen of the pair.

THE MEIER-SEASHORE ART JUDGMENT TEST RECORD SHEET

Directions:

In the accompanying booklet are pictures arranged in pairs, the two in each pair being very nearly alike. They differ only in *one* respect and you are told *what that is* in each case on pages 2, 3, and 4 of this blank.

You are to compare the two pictures in each pair, noting the unlike portion, and then decide which one is better (more pleasing, more artistic, more satisfying). Do not hurry. Study each pair carefully in turn.

Indicate your preference by drawing a circle about the "L" if you decide the left-hand picture is better, or around the "R" if you believe the right-hand one is more desirable.

L	R	121	Arrangement of the light, and picture
L	R	122	Borders and circular ornaments about bird
L	R	123	Location of pagoda in picture
L	R	124	Size of the middle ship
L	R	125	Position of the largest tree

FIGURE 14. ITEM 125, MEIER-SEASHORE ART JUDGMENT TEST.

The *McAdory Art Test* calls for reactions to a wide range of materials including furniture, clothing, rugs, etc. It may be given to individuals or to groups. The student responds to seventy-two multiple-answer treatments of the same form, of which twenty-four of the plates are in color. The testing equipment for this test is therefore naturally quite expensive.

255. **Measurement of Artistic Accomplishment.**[15] Numerous different types of tests and scales have been used in the measurement of achievement in the field of art. Only two of the most interesting and useful of these instruments are described here.

The *Kline-Carey Drawing Scales* consist of four separate scales each based upon fourteen specimens of drawing ranging from practically zero quality of representation to very high quality. The scales utilize the following types of subject-matter as themes: the house, a boy running, the tree (in silhouette), and the rabbit. An excellent feature of these scales is the series of verbal statements about the quality of each succeeding sample along with specific suggestions to the pupil for the improvement of his work.

A later series of scales by Kline and Carey measures elements of design and composition in free hand drawings. The four themes treated by these scales are illustrations, posters, structural design, and borders. Supervisors experienced in the measurement of art report a very definite field of service for these scales. They appear, however, to be somewhat limited by the fact that the unit of quality is not sufficiently fine to permit the expression of very minute differences. Possibly more extensive scales composed of many more specimens of smaller quality differences designed for use in a specific grade may be the solution of this difficulty.

Rating scales in art are subject to the same limitations encountered by all such scales in other subject-matter fields. The reliability of measurement resulting from their use depends upon (1) the representative nature of the specimens secured from the pupils for rating purposes, and (2) the

[15] All tests mentioned in this volume are listed, with publishers, in the *Appendix*, pages 585 to 595.

accuracy with which the judges are able to assign quality ratings. This second factor is unquestionably conditioned by the size of the unit of measurement used in establishing the scale. Some evidence on the use of rating scales in art reported in a recent article[16] indicates that "the use of a drawing scale decreases the inaccuracy of ratings to approximately one-half what it is when no scale is used."

TABLE 62

LEWERENZ TEST IN FUNDAMENTAL ABILITIES OF VISUAL ART

Test Number and Name	Time in Minutes	Number of Items	Form of Item
PART I			
1 Recognition of Proportion	10	14	M-C 4 answer
2 Originality of Line Drawing	20	10	Dot pattern
PART II			
3 Observation of Light and Shade	5	10	Picture completion
4 Knowledge of Subject-matter Vocabulary	20	60	Matching
5 Visual Memory of Proportion	5	1	Freehand drawing vase; bases and axis given
PART III			
6 Analysis of Problems in Cylindrical Perspective	5	7	Identification of error in drawing
7 Analysis of Problems in Parallel Perspective	5	5	Identification of error in drawing
8 Analysis of Problems in Angular Perspective	5	4	Identification of error in drawing
9 Recognition of Color	20	46	Identification, tints, shades

[16] Brooks, F. D., "The Relative Accuracy of Ratings Assigned With and Without the Use of Drawing Scales." *School and Society*, Vol. 27 : 518-20, April 28, 1928.

The *Lewerenz Test in Fundamental Abilities of Visual Art* is designed to measure specific aspects of accomplishment in art as a result of instruction. Nine tests arranged in three parts constitute the complete battery. Table 62 shows the list of tests, the time allowance for each test, and the type of testing technique used.

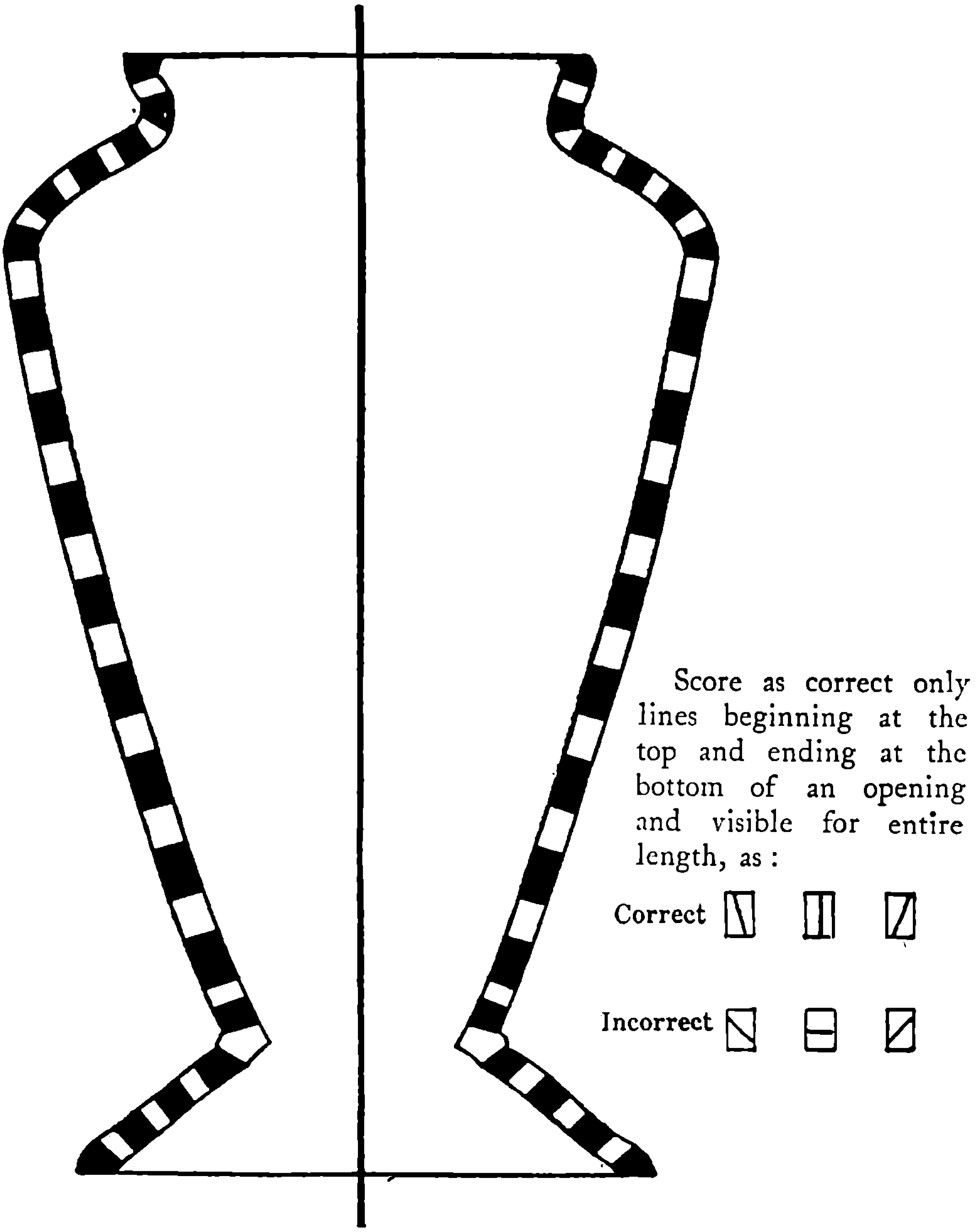
One of the most interesting scoring schemes the authors have seen is supplied for use with Test 5 (see Figure 15). Others of the tests are not so objectively scored but in spite of these limitations these tests are undoubtedly among the most useful standardized devices for the measurement of achievement in art.

256. **Applied Art.** Education will have failed in much of its social responsibility if it allows the child to leave the school without developing in him a rather definite love of the fine arts, even though it may be on a relatively low level. Not everyone can or should become musicians, painters or sculptors, but most everyone has the essential elements which make for a love and appreciation for the beautiful which he himself may be unable to produce. These elements, instruction in the fine arts should cultivate and develop. Furthermore, such instruction has a rather definite responsibility for making the arts function in real life in matters of personal adornment and in the planning and decorating of the home. The general cultural level of the population will be raised as this point is recognized and applied in instruction in the fine arts.

Discussion Exercises

1. In your opinion is there any reason to assume that achievement in the fine arts may not be objectively measured? Support your answer.
2. Is the measurement of musical talent or artistic talent any more difficult or basic than the measurement of aptitudes in any other complex field?
3. From your study of the available tests of musical talent, which of the aptitude tests discussed here seem to be most soundly grounded in critical research?
4. What is your chief criticism of the *Seashore Music Talent Tests? Kwalwasser-Dykema? Drake?*

Test 5. VISUAL MEMORY OF PROPORTION



Directions for Scoring : Place the scoring key over the paper to be scored so that the intersecting cross lines exactly coincide with those on the paper. The score is determined by the number of openings under which the outline drawn by the student falls.

FIGURE 15. SCORING KEY FOR LEWERENZ TESTS IN FUNDAMENTAL ABILITIES OF VISUAL ART[17]

[17] Reproduced with permission of publishers, Southern California School Book Depository, Los Angeles.

5. Which of the tests of musical achievement described in this chapter appear to overlap most widely the specified aims and outcomes of music instruction ?
6. Which one of the tests of musical accomplishment impresses you as being most suitable for classroom use ? Why ?
7. What similarities in the basic problems of measurement do you see in the fields of music and art ?
8. What are the major classes of general outcomes in art instruction ?
9. Which of the art tests described here seem to most adequately measure the major features of accomplishment in art ?
10. What is your chief criticism of the *Meier-Seashore Art Judgment Test ?* The *Lewerenz Test in Fundamental Abilities of Art ?*
11. Would you be willing to use the *Kline-Carey Drawing Scales* without some preliminary training in the rating of drawing specimens ? How would you go about securing such training or experience ?
12. Suggest a corrective program for certain art abilities revealed by the *Lewerenz Tests.*

Selected References

Brooks, Fowler D., "The Relative Accuracy of Ratings Assigned With and Without the Use of Drawing Scales." *School and Society,* Vol. 27 :518-20, April 28, 1928.

Dykema, P., *Tests and Measurements in Music Education.* Proceedings of Music Supervisors' National Conference, 1925.

Fifth and Sixth Yearbooks of the Department of Superintendence. Chapters on Music and Art. Washington, D. C. : National Education Association, 1927 and 1928.

Goodenough, Florence L., *Measurement of Intelligence by Drawings.* Yonkers, New York : World Book Company, 1926.

Judd, C. H., *Psychology of Secondary Education,* Chapter XIII. Boston : Ginn and Company, 1927.

Klar, Walter H., Winslow, Leon L., Kirby, C. Valentine, *Art Education in Principle and Practice.* Springfield, Mass. : Milton Bradley Company, 1933.

Kline, Linus W., and Carey, Gertrude L., *The Kline-Carey Measuring Scale for Freehand Drawing.* Johns Hopkins University Studies in Education, No. 5. Baltimore : Johns Hopkins Press, 1922.

Kwalwasser, J., *Tests and Measurements in Music.* Boston : C. C. Birchard and Company, 1927.

Kwalwasser, J., *Kwalwasser Test of Musical Information and Appreciation.* Iowa City : Bureau of Educational Research and Service, Extension Division, University of Iowa, 1927.

Kwalwasser, J., *Third Yearbook, Department of Superintendence.* "Music," report as chairman. Washington, D. C. : National Education Association.

Kwalwasser, J., Ruch, G. M., *Kwalwasser-Ruch Musical Achievement Tests.* Iowa City : Bureau of Educational Research and Service, Extension Division, University of Iowa, 1924.

Lewerenz, Alfred S., *Tests of Fundamental Abilities of Visual Art.* Los Angeles : Southern California School Book Depository, 1927.

Lewerenz, Alfred S., "Sex Differences on Ability Tests in Art." *Journal of Educational Psychology,* Vol. 19 :629-35, December, 1928.

Manuel, H. T., *Talent in Drawing.* School and Home Education Monographs, No. 3. Bloomington, Illinois : Public School Publishing Company, 1919.

McAdory, Margaret, *The Construction and Validation of an Art Test.* Contributions to Education, No. 383. New York : Teachers College, Columbia University, 1929.

Meier, Norman C., *Aesthetic Judgment as a Measure of Art Talent.* University of Iowa Series on Aims and Progress of Research, Vol. 1, No. 19. Iowa City : University of Iowa, 1926.

Meier, Norman C., and Seashore, C. E., *The Meier-Seashore Art Judgment Test.* Iowa City : Bureau of Educational Research and Service, Extension Division, University of Iowa, 1929.

Mursell, J. L., and Glenn, M., *The Psychology of School Music Teaching.* Boston : Silver, Burdett and Company, 1931.

Nyquist, F. V., *Art Education in Elementary Schools.* Baltimore : Warwick and York, 1929.

Pierce, Anne E., and Hilpert, Robert S., *Instruction in Music and Art, Monograph No. 25, National Survey of Secondary Education.* Washington, D. C. : United States Office of Education, 1932.

Seashore, C. E., *The Measurement of Musical Talent.* Boston : Silver, Burdett and Company, 1919.

Speer, Robert K., *Measurement of Appreciation in Poetry, Prose, and Art.* New York : Teachers College Contributions to Education, No. 362, Columbia University, 1929.

Stanton, H., *Prognosis of Musical Achievement,* University of Rochester, New York, 1929.

Uhl, W. L., and Others, *The Supervision of Secondary School Subjects.* Chapters XI and XII. New York : D. Appleton and Company, 1929.

Whitford, W. G., *An Introduction to Art Education.* New York : D. Appleton and Company, 1931.

CHAPTER XX

HEALTH AND PHYSICAL EDUCATION

This chapter is a summary of the following aspects of measurement and corrective instruction in the field of health and physical education:

a. Place of health and physical education in the general educational program.
b. Objectives of instruction in health and physical education.
c. Measurable qualities in health and physical education.
d. Preventive and corrective work in health and physical education.

I. Health and Physical Education in the Educational Program

257. **Educational Significance of Health Education.** This nation's greatest resource is the health and vitality of its people. The pages of history make this clear. In the interest of national welfare, therefore, all educational agencies must co-operate in giving systematic attention to safeguarding health. Intelligent programs of general health education are required to make each individual more completely aware of his own and the community's health problems, to give him a more adequate basis for solving these problems, and to train him to co-operate whole-heartedly in their solution.

The last quarter of a century has seen a rapid growth of interest in health education. This may be traced in part to the fact that educators have recognized health instruction and physical education as major educational outcomes. Upon the schools rests a large part of the burden of producing a well-informed generation that will insist upon adequate health protection. The evidence of increased illness among the unemployed and those immediately affected by

their unemployment sounds a serious warning. It must be recognized that in addition to the well-known rights of "life, liberty and the pursuit of happiness," health is also an inalienable right of the American people. Health protection through general enlightenment and readily available medical service is a necessity of life in almost the same degree as are food, clothing, and shelter.

258. Objectives of Instruction in Health. Health education has been defined as "the sum of experiences in school and elsewhere, which favorably influence habits, attitudes, and knowledge related to individual, racial, and community health."[1] This same general idea has been expressed by Burton[2] as follows: "Health education in the schools may be defined as the procedure employed in supervision and teaching to help children live healthfully." In contrast, physical education has usually been regarded as merely exercise and disciplinary drill. For schools, physical education includes the sum of the individual's physical activities, selected and organized according to some plan and conducted according to defined outcomes. The selection of activities is guided by the principle that only those possessing real educative value will be chosen. Each physical activity will then be carried on to achieve the mental, moral, social and physical values inherent in the activities themselves.

The purpose of health instruction is to secure a wholesome physical environment in which to develop child health, to teach habits, skills, facts, and attitudes favorable to health. The specific educational objectives of health education have been stated concisely as follows:

1. To instruct children and youth so that they may conserve and improve their own health.
2. To establish in them the habits and principles of living which throughout their school life, and in later years, will assure that abundant vigor and vitality which provide the basis for the greatest possible happiness and service in personal, family, and community life.
3. To influence parents and other adults through the health education program for children, to better habits and attitudes, so that the

[1] Wood, T. D., and Brownell, C. L., *Source Book in Health and Physical Education.* New York: The Macmillan Company, 1925.

[2] Burton, W. H., and others, *The Supervision of Elementary Subjects.* New York: D. Appleton and Company, 1929.

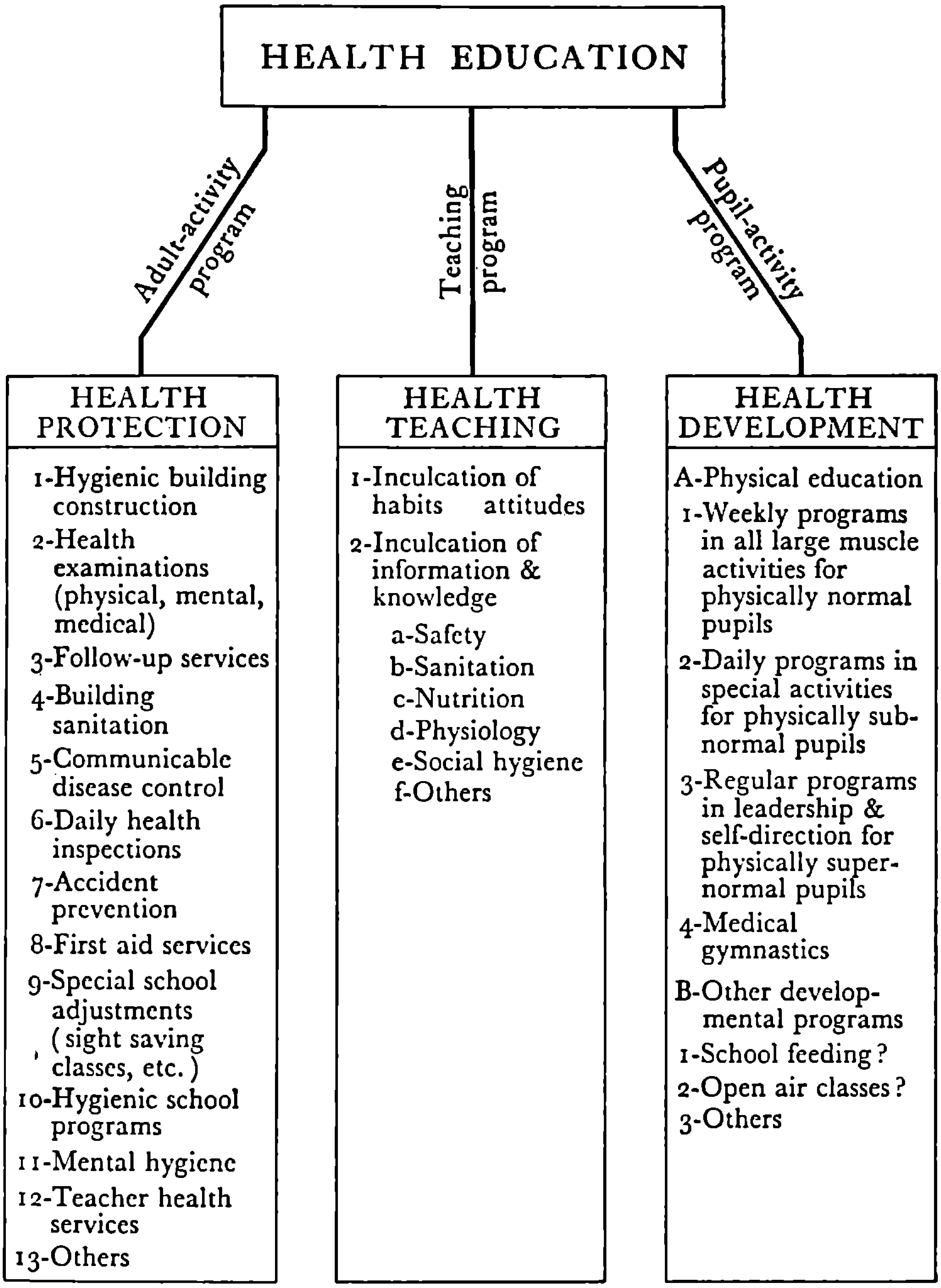


FIGURE 16. A PROGRAM TO PROTECT AND CONSERVE HEALTH [3]

[3] *Health Education — A Manual of Information for Teachers.* University of the State of New York Bulletin. Albany, N. Y.

school may become an effective agency for the promotion of the social aspects of health education in the family and community as well as in the school itself.

4. To improve the individual and community life of the future ; to insure a better second generation, and a still better third generation ; a healthier and fitter nation and race.[4]

In the light of these objectives, health is not merely a matter of sanitary environment, freedom from physical defect, or a matter of abundant health knowledge. It is *a way of living*. To make sure that the way will be the best way, supervision of the environment, correction of defects, and health instruction are necessary. Something of the detail of an effective health education is shown in Figure 16.

259. **Objectives of Physical Education.** The purpose of physical education is to offer an opportunity for the pupil to act in situations which are physically wholesome, mentally stimulating and satisfying, and socially sound. The objectives of physical education may be stated as follows :

1. To develop the organic system of the individual through suitable physical activities.
2. To develop the neuro-muscular system in general, and particularly in relation to control over certain fundamental skills.
3. To develop certain attitudes toward physical activity generally and toward play particularly.
4. To develop standards of conduct.

Each pupil is equipped by nature to engage in many natural physical activities. What he requires most are opportunities for engaging in such activities. Adequate opportunities for physical education will include physical facilities, leadership, guidance, and supervision so that these activities may go forward for the maximum satisfaction and achievement of all participants. From the school's point of view, a well planned program is one which will result in the practical conduct of these activities directed to secure wholesome physical effects, permanent attitudes favoring play and recreation, and desirable social and moral conduct. The pro-

[4] *Health Education : A Program for Public Schools and Teacher Training Institutions.* A Report of the Joint Committee on Health Problems in Education of the National Education Association and the American Medical Association with the co-operation of the Technical Committee of Twenty-Seven.

gram of physical education work may be visualized by studying Figure 17.

260. Measurable Qualities in Health and Physical Education. While the objectives for health and physical education are useful for centering attention upon desirable outcomes of instruction they do not adequately define the content of the

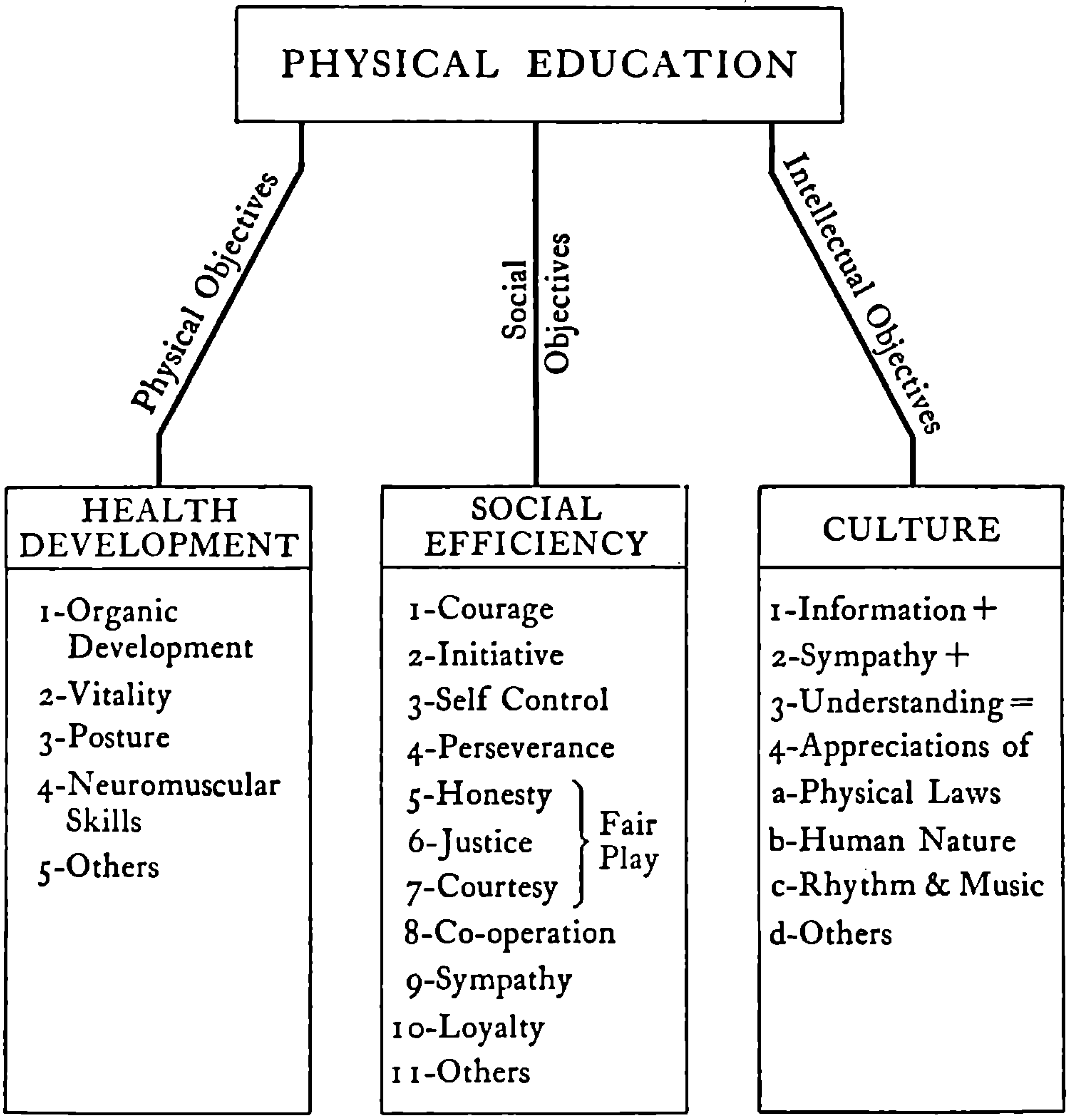


FIGURE 17. A PROGRAM TO PROTECT AND IMPROVE HEALTH AND SOCIAL EFFICIENCY [5]

courses nor specify the items to be incorporated in measuring instruments in this field.

The two most important outcomes of instruction in these subjects are (1) the development of adequate health habits, and (2) the creation of proper attitudes toward healthful living. These habits and attitudes are difficult to meas-

[5] *Ibid.*

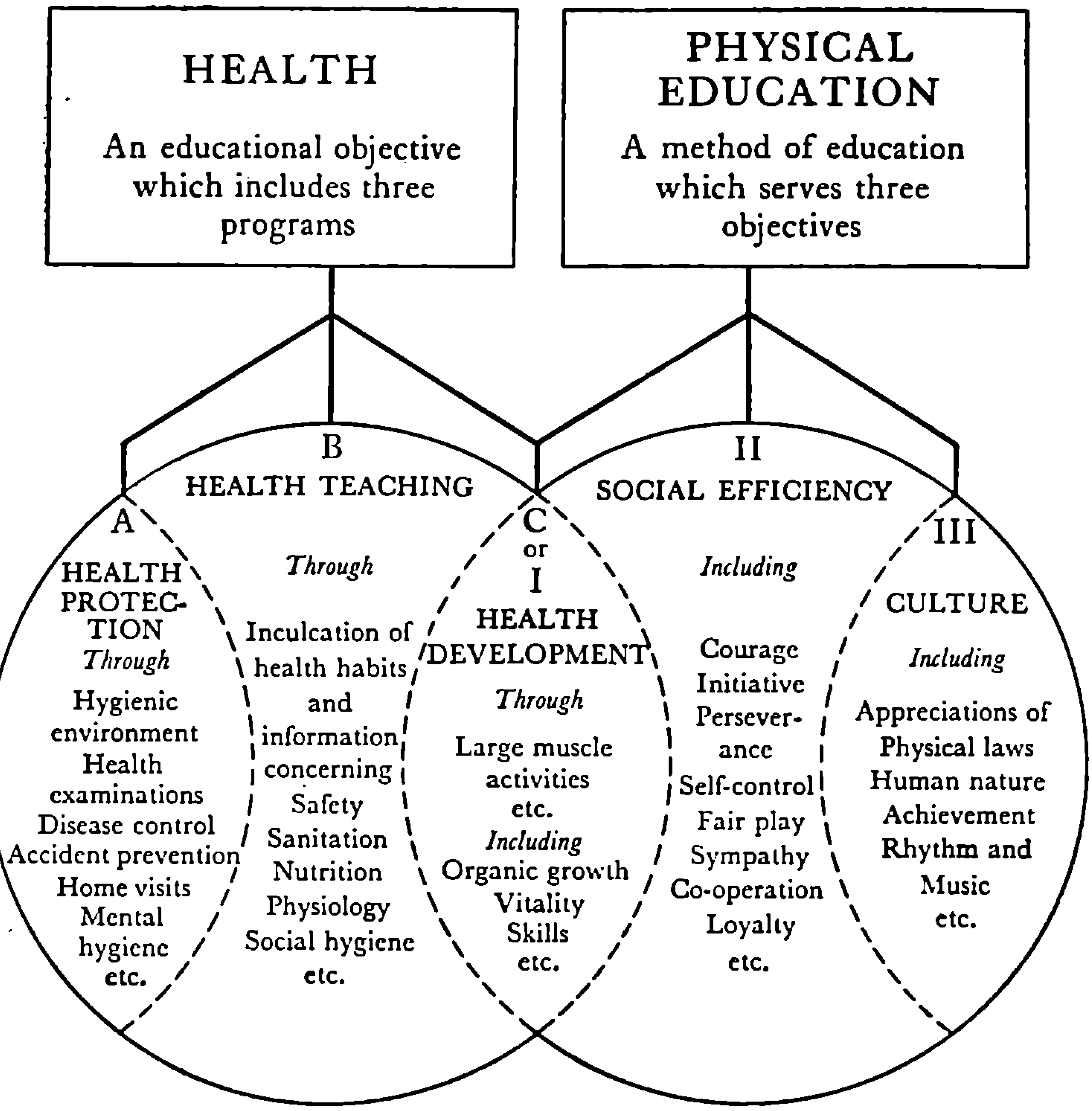


Fig. 18. Relation of Health and Physical Education [6]

ure, but among the elements basic to their development and capable of reasonably objective measurement are the following:

1. Health knowledge
2. Health habits
3. Physical growth
4. Physical capacity
5. Motor abilities
6. General achievement
7. Athletic proficiency

This catalogue of measurable qualities makes it apparent that health education and physical education are so highly related that as a part of the educational program they may for practical purposes be regarded as one. Obviously, a close

[6] *Ibid.*

relationship of health work, physical education, intramural, and interscholastic athletics should be obtained in the secondary schools. The work of all four fields should be coordinated and administered under one head, with the single objective of improving individual and community health. This inter-relationship is well shown in Figure 18.

II. Measurement of Information on Health

Many health tests are of more value to the physician and school nurse than to the teacher. They include such procedures as testing for acuity of hearing, acuity of vision and the like. The techniques and the equipment used for such purposes are entirely too elaborate for the classroom teacher. Health tests which are of most interest and value to the teacher are those whose function is to reveal, as the basis for the formation of desirable health habits and attitudes, the extent of health knowledge possessed by the pupil.

261. **Health Knowledge Tests.**[7] The *Gates-Strang Health Knowledge Test* is designed to measure a pupil's knowledge of many phases of healthful living. The test is based upon a careful analysis of facts found in twenty selected courses in health and in fourteen widely used textbooks on health. From this analysis, 520 items were found to be common to three or more of these sources. These items sampled the following thirty phases of health education:

Food	98
Disease Prevention	78
Physiology	33
Exercise and Posture	30
Cleanliness	27
Fresh Air	21
Mental Hygiene	18
Care of the Eyes	17
Safety	15
Defects, including Malnutrition	14
Clothing	14
Care of the Teeth	14
Water	13

[7] All tests mentioned in this volume are listed with publishers in the *Appendix*, pages 585-595.

Industrial Hygiene	12
Values of Health	12
First Aid	11
Elimination of Body Wastes	10
Care of the Hair	10
Rest and Sleep	10
Disposal of Garbage and Waste	10
Child Care	9
Names of Scientists and Miscellaneous Items	8
Alcohol	7
Care of Ears, Nose, and Throat	6
Care of the Feet	6
Patent Medicines and Drugs	5
Tobacco	4
Health Laws	3
Public Health Administration	3
Health Organizations	2

The items of these thirty groups were incorporated into a series of standardized tests containing 64 test items. The items are of the recognition type with a choice of five alternative answers. The test may be used in grades 3 to 12. Norms for grades 3 to 8 are available. The time required for giving the test is 40 minutes.

While the *Health Education Tests* of the *American Child Health Association* are much too easy for high school use they are without question very suggestive instruments. Five tests, all of the paper and pencil type, comprise the battery. The authors recommend that when all five tests are used they be given in the order in which they appear in Table 63.

TABLE 63

HEALTH EDUCATION TESTS

Test	Testing Time in Minutes	Average Time in Minutes for Scoring 40 Papers
Five Rules	4	20
Story	30	100
Time	8	45
Matching	30	100
True-False	20	300

The *Five Rules Test* comprises twenty rules arranged in two columns. In each column the pupil is to indicate the five rules which he would most like to follow for one month. The *Study Test* contains seven stories each of approximately three hundred words in length. In this test the pupil is to underline all statements that deal with things good for the health, and mark out all statements of things which are bad for the health. The *Time Test* presents twenty-two statements with directions for marking in special ways things which should be attended to at least once a day, once a week or once a month, once or twice a year, etc. The *Matching Test* presents sixty-three items of health information in nine groups for convenience in scoring. The *True-False Test* contains ninety-two items designed to sample other aspects of health information.

The *Orleans-Sealy Health Test* comprises one section of the *Public School Achievement Tests.* These tests are standardized for use in grades four to eight inclusive. The four-page test itself is made up of four parts presenting 42 items in true-false form, 13 completion items, 11 multiple-choice items, and 18 matching items. The total testing time is twenty minutes.

The *Wood-Lerrigo Health Scale* is published as a teacher's manual under the title of *Health Behavior.* This manual contains six scales the purpose of which is to measure the results of health instruction. The items of the scales may be used for the organization of tests to measure the pupil's knowledge of health information, habits, and practices. The standards for the scales are based upon a study made by the authors for the *Joint Committee on Health Problems in Education* of the *American Medical Association* and the *National Education Association.*

III. Measurement of Physical Growth and Strength

The early emphasis in the measurement of physical growth was upon symmetry and size. It was not until about 1880 that the work of Hitchcock and Sargent centered attention upon strength as revealed by actual work done by the

individual. Sargent, while still a medical student at Yale, began working on the idea of strength as a measure of physical development and capacity. The measurement of strength received considerable impetus with the development of the spirometer and the dynamometer.

Recently strength tests have declined in popularity. Observation has revealed the fact that frequently large men are not always the men of high endurance or of greatest strength, nor are they always able to use their strength to best advantage. These facts led to new developments in physical measurements based upon the idea that what is really sought for in physical education is not merely measures of inherent strength, or size, or symmetry, but of effective ability to use muscle power in handling the body in the various activities of everyday life.

It is not to be assumed, however, that anthropometric measurements and strength tests have been entirely discarded. Both contained many elements of value for measurement in physical education. The factors of age, height and weight are important regardless of the type of test used. Furthermore, the idea of physical capacity inherent in the older strength tests is also of value in present measurements in pointing out to each person how he compares with others of his age and size. It is also of value in assisting the physical education instructor in prescribing to each person the physical exercises most suitable for him. Rogers[8] has recently revived certain of the older strength tests and has shown that many of them are valid as measures of general athletic ability. In their improved form they are used to classify boys for purposes of competition in the large-muscle exercises and athletic events.

The remainder of this section is devoted to brief descriptions of several tests available for use in measuring physical growth, capacity, motor ability, achievement and athletic proficiency.

262. **Weight-Height-Age Tables.** The Baldwin-Wood *Weight-Height-Age Tables* are presented here as typical of

[8] Rogers, F. R., *Tests and Measurement Programs in the Redirection of Physical Education.*

a number of such comparative summaries. These tables show the measurements of American-born children in inches and pounds. Age is taken to the nearest birthday. Weight is taken to the nearest pound. In these tables a certain percentage of net weight has been included to allow for clothing, except shoes and sweaters, which are removed when weight is taken. For weights from 35 to 63 pounds, 3.5 per cent of the net weight has been included. For weights of 64 pounds or more, 4 per cent of the net weight has been included.

After the measures of height and weight of each pupil have been accurately taken, they are interpreted by reference to the tables as follows:

(1) Locate the column at the top of the table representing the pupil's age.

(2) Follow this column down until even with the horizontal line which represents the child's height. The figure at that place in the column that cuts the horizontal line represents the normal weight for the age and height of the pupil measured.

(3) Compare the weight and height of the pupil measured with the weight given in the table.

(4) Determine the difference between these two figures. This obtained difference is taken to indicate the physical condition of the pupil. If the pupil's weight exceeds the standard given in the table, he is marked plus. If his weight is below the standard given, he is marked minus. Exactly what per cent of underweight or overweight is to be considered serious is not fully agreed upon. Generally a figure which deviates from the given standard by 10% is indicative of a weight condition which should be carefully diagnosed.

While growth is an excellent indication of good health, too much importance should not be placed upon any one of its particular phases. For example, emphasis upon weight alone does not represent a reliable measure of health. Overweight or under-weight in itself may mean little. Many factors influence weight, such as age, height, and size and form of bony structure. Even *weight-height-age* tables,

Weight-Height-Age Tables in English Units of American-Born Boys (Clothed) of School Age

By Bird T. Baldwin, Iowa Child Welfare Research Station, State University of Iowa, and Thomas D. Wood, Teachers College, Columbia University

Mean Weight of Boys According to Height and Age

Height Inches	Age, Years 6	7	8	9	10	11	12	13	14	15	16	17	18	Height Inches
	Weight, Pounds													
37.5—38.4	34*	34*	34*	...	...	...	...	...	...	...	...	...	...	37.5—38.4
38.5—39.4	35*	35*	35*	...	...	...	...	...	...	...	...	...	...	38.5—39.4
39.5—40.4	36*	36*	36*	...	...	...	...	...	...	...	...	...	...	39.5—40.4
40.5—41.4	38	38*	38*	...	...	...	...	...	...	...	...	...	...	40.5—41.4
41.5—42.4	39	39*	39*	40*	...	...	...	...	...	...	...	...	...	41.5—42.4
42.5—43.4	41	41*	41*	42*	...	...	...	...	...	...	...	...	...	42.5—43.4
43.5—44.4	44	44	44*	44*	45*	...	...	...	...	...	...	...	...	43.5—44.4
44.5—45.4	46	46	46*	46*	47*	...	...	...	...	...	...	...	...	44.5—45.4
45.5—46.4	48	48	48	48*	49*	49*	...	...	...	...	...	...	...	45.5—46.4
46.5—47.4	50	50	50	50*	50*	51*	51*	...	...	...	...	...	...	46.5—47.4
47.5—48.4	52	53	53	53	53*	54*	54*	...	...	...	...	...	...	47.5—48.4
48.5—49.4	55	55	55	55	55	55*	56*	...	...	...	...	...	...	48.5—49.4
49.5—50.4	57*	58	58	58	58	58*	58*	59*	59*	60*	...	...	...	49.5—50.4
50.5—51.4	60*	61	61	61	61	61	61*	62*	62*	63*	...	...	...	50.5—51.4
51.5—52.4	63*	63	64	64	64	64	64	64*	65*	66*	...	...	...	51.5—52.4
52.5—53.4	65*	66*	67	67	67	67	68	68*	69*	70*	...	...	...	52.5—53.4
53.5—54.4	68*	69*	70	70	70	70	71	71	72*	73*	...	...	...	53.5—54.4
54.5—55.4	...	71*	72*	72	73	73	74	74	74*	76*	77*	...	...	54.5—55.4
55.5—56.4	...	74*	75*	76	77	77	77	78	78	80*	81*	...	...	55.5—56.4
56.5—57.4	...	...	78*	79*	80	81	81	82	83	83*	84*	...	...	56.5—57.4
57.5—58.4	...	...	82*	83*	84	84	85	85	86	87	88*	...	...	57.5—58.4
58.5—59.4	...	...	...	86*	87	88	89	89	90	90	90	...	...	58.5—59.4
59.5—60.4	...	...	...	90*	91*	92	92	93	94	95	96	97*	...	59.5—60.4
60.5—61.4	...	...	...	...	94*	95	96	97	99	100	103	106*	110*	60.5—61.4
61.5—62.4	...	...	...	...	97*	100*	101	102	103	104	107	111	116*	61.5—62.4
62.5—63.4	...	...	...	...	102*	105*	106	107	108	110	113	118	123	62.5—63.4
63.5—64.4	...	...	...	...	106*	108*	109	111	113	115	117	121	126	63.5—64.4
64.5—65.4	...	...	...	...	110*	112*	114*	117	118	120	122	127	131	64.5—65.4
65.5—66.4	...	...	...	...	113*	115*	117*	119	122	125	128	132	136	65.5—66.4
66.5—67.4	...	...	...	...	...	...	...	124*	128	130	134	136	139	66.5—67.4
67.5—68.4	...	...	...	...	...	...	...	...	134	134	137	141	143	67.5—68.4
68.5—69.4	...	...	...	...	...	...	...	...	137	139	143	146	149	68.5—69.4
69.5—70.4	...	...	...	...	...	...	...	...	143	144	145	148	151	69.5—70.4
70.5—71.4	...	...	...	...	...	...	...	...	148*	150	151	152	154	70.5—71.4
71.5—72.4	...	...	...	...	...	...	...	...	...	153	155	156	158	71.5—72.4
72.5—73.4	...	...	...	...	...	...	...	...	...	157*	160	162	164	72.5—73.4
73.5—74.4	...	...	...	...	...	...	...	...	...	160*	164	168	170	73.5—74.4

Weight-Height-Age Tables in English Units of American-Born Girls (Clothed) of School Age

By Bird T. Baldwin, Iowa Child Welfare Research Station, State University of Iowa, and Thomas D. Wood, Teachers College, Columbia University

Mean Weight of Girls According to Height and Age

Height Inches	Age, Years													Height Inches
	6	7	8	9	10	11	12	13	14	15	16	17	18	
	Weight, Pounds													
37.5—38.4	33	33*	...	...	...	...	...	...	...	...	...	...	...	37.5—38.4
38.5—39.4	34	35*	...	...	...	...	...	...	...	...	...	...	...	38.5—39.4
39.5—40.4	36	36*	...	...	...	...	...	...	...	...	...	...	...	39.5—40.4
40.5—41.4	37*	37*	37*	...	...	...	...	...	...	...	...	...	...	40.5—41.4
41.5—42.4	39	39	39*	40*	...	...	...	...	...	...	...	...	...	41.5—42.4
42.5—43.4	41	41	41*	41*	...	...	...	...	...	...	...	...	...	42.5—43.4
43.5—44.4	42	42	42*	43*	44*	44*	44*	...	...	...	...	...	...	43.5—44.4
44.5—45.4	45	45	45	45*	46*	46*	46*	...	...	...	...	...	...	44.5—45.4
45.4—46.4	47	47	48	48*	48*	49*	49*	...	...	...	...	...	...	45.5—46.4
46.5—47.4	50	50	50	50	50*	51*	52*	...	...	...	...	...	...	46.5—47.4
47.5—48.4	52	52	52	52	53*	53*	54*	...	...	...	...	...	...	47.5—48.4
48.5—49.4	54	54	55	55	56	56*	58*	58*	58*	...	...	...	...	48.5—49.4
49.5—50.4	56*	56	57	58	59	61	62*	63*	64*	...	...	...	...	49.5—50.4
50.5—51.4	59*	59	60	61	61	63	65	66*	69*	...	...	...	...	50.5—51.4
51.5—52.4	62*	63*	64	64	64	65	67	69*	71*	...	...	...	...	51.5—52.4
52.5—53.4	65*	66*	67	67	68	68	69	71*	73*	...	...	...	...	52.5—53.4
53.5—54.4	...	69*	69	70	70	71	71	73*	75*	...	...	...	...	53.5—54.4
54.5—55.4	...	72*	72*	74	74	74	75	77	78*	79*	...	...	...	54.5—55.4
55.5—56.4	...	75*	76*	76	78	78	79	81	83*	86*	89*	...	...	55.5—56.4
56.5—57.4	...	...	79*	80*	82	82	82	84	88	92*	96*	100*	...	56.5—57.4
57.5—58.4	...	...	...	84*	84	86	86	88	93	96*	101*	103*	...	57.5—58.4
58.5—59.4	...	...	...	87*	87	90	90	92	96	100	103*	104*	...	58.5—59.4
59.5—60.4	...	...	...	90*	91*	95	95	97	101	105	108	109	111*	59.5—60.4
60.5—61.4	...	...	...	...	96*	99	100	101	105	108	112	113	116	60.5—61.4
61.5—62.4	...	...	...	...	101*	104*	105	106	109	113	115	117	118	61.5—62.4
62.5—63.4	...	...	...	...	...	108*	110	110	112	116	117	119	120	62.5—63.4
63.5—64.4	...	...	...	...	...	112*	114*	115	117	119	120	122	123	63.5—64.4
64.5—65.4	...	...	...	...	...	116*	118*	120	121	122	123	125	126	64.5—65.4
65.5—66.4	...	...	...	...	...	...	122*	124	124	125	128	129	130	65.5—66.4
66.5—67.4	...	...	...	...	...	...	126*	128*	130	131	133	133	135	66.5—67.4
67.5—68.4	...	...	...	...	...	...	130*	131*	133	135	136	138	138	67.5—68.4
68.5—69.4	...	...	...	...	...	...	...	134*	135*	137*	138*	140*	142*	68.5—69.4
69.5—70.4	...	...	...	...	...	...	...	136*	136*	138*	140*	142*	144*	69.5—70.4
70.5—71.4	...	...	...	...	...	...	...	138*	138*	140*	142*	144*	145*	70.5—71.4

*The starred figures represent values based upon theoretical computations instead of on exact averages.
Age is taken at the nearest birthday; for example, a child is considered 6 years old from 5 years, 6 months to 6 years, 5 months, and 29 days.

although of tremendous value when used intelligently, cannot be relied upon entirely. However, gains should not be too seriously regarded in the light of a pupil's age and height standards until hereditary and all other health factors are determined. The data for each child with respect to his age, height, and weight must be interpreted in the light of all other available information concerning his health. The general practice of measuring for weight and health is valuable for its educational implications, because it centers the attention of children upon health and growth standards and encourages them to form desirable health habits.

263. **Physical Capacity Tests.** Physical fitness and proficiency are usually determined by capacity tests based upon a wide range of physical performances for which standards have been established. "Physical fitness connotes capacity for purposeful physical activity—the functions of locomotion, manipulation and transportation; of 'doing things with matter'; of voluntary conflict with the natural forces of gravitation and inertia."[9]

The *Rogers Strength Index* was devised to assist in grouping boys for purposes of team competition within the school and within class groups. According to Rogers, "the strength index, which is a score indicating the strength of the large voluntary muscles of the body, together with lung capacity, conforms to all the criteria of a useful measure of general athletic ability. First, and most important, it is a highly valid measure, being over two and one-half times as accurate as weight, and nearly twice as accurate as the best possible combination of age, height and weight. It is economical of time; boys can be given all tests at the rate of one boy per minute, using only a single adult tester with a few student assistants, and indices can be computed in fifteen to twenty seconds each. Boys are fascinated by the testing procedure, often asking for retests. Thus their wholehearted co-operation is secured without difficulty. The tests are well adapted to the age and sex of the subjects; they are easily and accurately scored in mathematical terms; they are objective, and the strength index is probably more

[9] Rogers, F. R., *Fundamental Administrative Measures in Physical Education.* The Pleiades Company, Newton, Massachusetts, 1932. P. 116.

reliable than any highly valid composite mental test available."[10]

The test consists of seven parts:

1. Number of cubic inches in lung capacity, measured with wet spirometer.
2 and 3. Strength of grip (of forearm) (flexion) of each hand, measured by grasping the elliptical manometer or hand-dynamometer, in pounds.
4. Strength of back (extension) measured by lifting with knees straight, using the back and leg dynamometer, in pounds.
5. Strength of legs (extension) measured by lifting, carrying the weight on the thighs, with the back and leg dynamometer.
6. Strength of arms (and shoulder girdle) (extension) by "push-ups" from straight-arm hang on the parallel bars, as many times as possible.
7. Strength of arms (and shoulder girdle) (flexion) by "chinning" or "pull-ups," using light, inch rings attached to overhead ladder.

The strength of the arms is calculated as follows:

$$\text{Pull-ups} + \text{push-ups} \times \left[\left(\frac{\text{weight}}{10} + (\text{height} - 60)\right]$$

The scores on these tests are combined into a composite score, a *Strength Index*. The reliability of this test, based on 644 students in junior and senior high schools, is between .94 and .98. The use of the table of strength norms[11] for various age and weight measurements permits the calculation of *Physical Fitness Indices.* Rogers defines the physical fitness index as "the achieved strength index of any individual divided by the normal strength index for the individual's age and weight."[12] After the division the index is multiplied by 100 to eliminate decimals. Scores above or below 100 indicate the percentage above or below the normal.

The *McCloy Measurement of General Motor Capacity* is designed to measure *innate motor capacity.* McCloy defines this term as "the inborn, hereditary potentialities for general motor performance. The word *general* in this definition is in contrast to specific abilities such as the ability to play basket-ball."

[10] Rogers, F. R., Op. Cit., p. 18.
[11] Rogers, F. R., Op. Cit., p. 155.
[12] Rogers, F. R., Op. Cit., p. 152.

For junior and senior high school boys there are four parts to the test; for girls, three.

1. *Classification index.*[13] (This test is omitted for girls.) This index is computed as follows:

20 (age in years) + 6 (height in inches) + (weight in pounds)

2. *Iowa Revision of the Brace Test.* The original Brace Test[14] consisted of 20 stunts given in four alternating batteries of five. The Iowa Revision consists of twenty-one stunts selected from the twenty and supplemented by new ones which are grouped in two batteries of five stunts each. The selection of tests differs for each group, as shown below.

THE STUNTS FOR EACH GROUPING IN SCHOOL ARE AS FOLLOWS:

Junior High School

Boys		*Girls*	
First half	*Second half*	*First half*	*Second half*
1	2	2	1
14	3	12	13
13	12	15	11
19	16	19	16
6	17	17	20

Senior High School

Boys		*Girls*	
First half	*Second half*	*First half*	*Second half*
1	3	3	3
11	14	11	18
16	15	7	16
5	17	17	9
20	21	19	20

The procedure in giving the test is to divide the group to be tested into two sections. One section will take the first half of the test and section two, the second half. The two groups then reverse the order for the remainder of the test.

13 Tables for computing this index may be found in McCloy, *Measurement of Athletic Power*, A. S. Barnes Co.

14 Brace, David K., *Measuring Motor Ability*, A. S. Barnes Co.

The tests are as follows:

Test	1.	(See Brace Test No. 13)[15]
Test	2.	Side leaning rest
Test	3.	(See Brace Test No. 14)
Test	4.	One knee balance
Test	5.	(See Brace No. 18)
Test	6.	(See Brace No. 8)
Test	7.	(See Brace No. 17)
Test	8.	(See Brace No. 7)
Test	9.	One knee ; head on floor
Test	10.	Hop backward
Test	11.	Forward hand kick
Test	12.	Full squat ; arm circles
Test	13.	Half turn jump ; left foot
Test	14.	(See Brace Test No. 5)
Test	15.	Side kick
Test	16.	(See Brace Test No. 16)
Test	17.	Russian Dance
Test	18.	(See Brace Test No. 15)
Test	19.	The top
Test	20.	Single squat balance
Test	21.	(See Brace Test No. 10)

3. *The Sargent Jump.* "This test consists in jumping straight upward so as to project the crown of the head as high as possible. The measurement is from the original standing height of the head to the highest point reached in the jump."

4. *The Burpee Test.* "The pupil stands erect in the ordinary standing position. At the command to start, he performs the following exercises as fast as possible and continues for ten seconds:

"*a.* Bend forward, bending knees as much as is desired, and place the hands on the floor in front of the feet ;
"*b.* Extend the legs backward to the position of the front leaning rest ;
"*c.* Return to (*a*) ; and
"*d.* Return to the erect position."

The above tests are scored using tables provided for that purpose. Standards for various classification indices (boys) or ages (girls) are provided. The standard for an individual pupil may be divided into the actual test score to obtain

[15] See list of tests in the Brace Scale on pages 510-511.

a motor quotient. Such a motor quotient is similar to the intelligence quotient in the mental field while the general motor capacity score is analogous to the gross raw score secured by use of an intelligence test. The motor quotient will reveal how the pupil ranks when compared to others of equal size and maturity. A quotient of 100 represents average capacity. The motor quotient is useful for sectioning pupils into groups and for assisting in marking and grading.

264. Scales of Motor Ability. The changing emphasis in physical education from gymnastics and calisthenics to games and sports is responsible for the development of tests of motor capacity and physical achievement. The analysis of physical ability into its component elements such as speed, accuracy, co-ordination, agility, balance, and endurance has resulted in the search for games or sports that contain these elements, and the preparation of tests capable of revealing the degree of skill possessed by each pupil in these elements. This naturally has centered attention upon the problem of evaluating and scoring the abilities revealed, also upon the problem of classifying pupils by age, weight, height, and school grade, in order to assure that game competition may be placed upon a fair competitive basis. Fundamental to all such testing is the question of the motor ability possessed by each pupil. An inventory of the motor abilities possessed and the degree of their development is obviously of great importance for the intelligent planning of the physical education program.

Brace's Scale of Motor Ability Tests is designed to measure native motor ability in pupils for ages 8 to 18 inclusive. The scale consists of 20 events in the nature of stunts which are easy to administer and to score. All the tests are scored on the basis of success or failure.

The stunts consist of the following:

1. Walking in a straight line.
2. Stand, jump into the air and clap both feet together once, and land with the feet apart any distance.
3. Trunk raising with feet on the floor.
4. Kneel on both knees and recover with arms folded behind the back.
5. Floor push up three times in succession.

6. Floor spring with hands clasping toes, to feet apart and arms sideward.
7. Full turn left in the air and land without losing balance.
8. As in 2 except feet clapped together twice.
9. Bending and touching left knee to floor while grasping left foot behind right knee.
10. Jumping through loop formed by grasping one toe with opposite hand.
11. Jump into air and slap both heels with hands behind back.
12. High kick so that toes come at least level with shoulders.
13. Forward bend to floor with both hands and head touching and right leg extended backward.
14. Full bend with knees, arms between knees, and around ankles, holding for five seconds.
15. As in 7 turning right.
16. Jumping to feet from kneeling position.
17. Cross legs and arms. Sit and regain position.
18. Hold the following position for ten seconds with eyes closed ; hands on hips and bottom of right foot against left knee.
19. Frog stand for five seconds.
20. One knee dip with foot extended forward and recover position.

The author makes the following statements with reference to the use of these motor ability tests :[16]

1. Motor ability scale scores may be used as the basis for determining an accomplishment quotient for the activities of physical education.
2. The scale of tests should form a valuable element in a scheme of classification of pupils for class work in physical education. This should be especially valuable in programs which favour the individual development of pupils.
3. Diagnosis of special performance disabilities may be assisted by a study of individual reactions to these tests. . .
4. Experimental studies in physical education which require the equating of groups of pupils have been greatly handicapped by the lack of standardized tests of general motor ability. . .
5. Finally, the scale of tests should stimulate other scientific efforts in the field of tests and measurements in physical education. Leads thus far established should assist other research on the nature of motor ability.

The McCloy General Motor Ability Test. This test consists of four track and field events in addition to the *Chinning Test.* These four events are :

[16] Brace, D. K., *Measuring Motor Ability.* New York : A. S. Barnes and Company, 1927.

a. The sixty yard dash.
b. The running high jump (omitted in testing girls).
c. The standing broad jump.
d. Either the shotput or the basket-ball throw.

Carefully formulated rules for the administration and scoring of these five tests are provided. According to the author of the test, "the *general motor ability test*, while limited to track and field athletics and to chinning, has a wider applicability than this would indicate. There is evidence that chinning measures total strength to within a correlation of approximately .90, while track and field athletics correlate very high with basket-ball and other team games. It must be recognized, however, that the narrow scope of specialization of this test limits its usefulness to that of a test of *general* motor ability."

The score on this series of tests is referred to as the *General Motor Ability Score* and according to McCloy "corresponds in the physical measurement field to the raw score of a composite achievement test in the mental measurement field. This score also is relative to the size and maturity of the person measured, and is useful only when combined with the General Motor Capacity Score[17] into the General Motor Ability Quotient." The *Motor Quotient* is a ratio of the General Motor Capacity Score and the General Motor Ability Score. This quotient is useful in revealing how good each subject tested is for his capacity or potentiality and indicates how he ranks when compared with others of equal size and maturity.

The author has placed the standard of the test at two standard deviations above the mean performance, rather than at the average. For this he gives two reasons. "The *limitation* of two standard deviations gives a performance standard that is not above the realm of possibility for most individuals in this test, and is, on the other hand, not so high as to be an encouragement to excessive strain and training. The second reason is that this quotient can be thought of by the pupil as a percentage score, with 100% as the desirable

[17] See discussion of the *General Motor Capacity Test* (McCloy) in the preceding section.

level. This is, of course, purely arbitrary and is not based upon any other fundamentally scientific principles."

The primary purpose of such tests as the two described in this section is that of classifying pupils and of furnishing a basis upon which to evaluate achievement.

265. Physical Achievement Tests.[18] After measures of physical capacity and the relative motor abilities of an individual have been secured as the basis for physical instruction, measures of achievement are needed. Measurement of achievement in physical education has done much to set up standards of achievement, to stimulate interest in all types of physical endeavor, and to raise the general physical efficiency of boys and girls.

The purpose of the *Philadelphia Public School Age Aim Charts* is to stimulate interest in physical skills and further participation in athletic activities.

The factors involved in the preparation of these charts are as follows:

1. The setting up of standards of achievement for given ages.
2. The awarding of a certificate of attainment for those who have passed certain year standards.
3. The establishment of a *physical quotient* for each boy and girl in the high school system.
4. The establishment of a scoring scheme for girls.

To be eligible for the events and athletic ability tests for the high school division a pupil must:

1. Be under 14 years of age for Test I, and under seventeen years of age for Tests II and III.
2. Have nearly or completely attained the standards of the test to be taken.
3. Have passed the age aims.

TEST I (FIFTEEN YEAR STANDARDS)

	Boys	Girls
1. Standing Broad Jump..	5 feet 10 inches	4 feet 10 inches
2. Ball Throw (overhead).	33 feet	28 feet
3. 50 Yard Dash........	7 3/5 seconds	8 2/5 seconds
4. Ball Throw (free style)	45 feet	35 feet

[18] All tests mentioned in this chapter are listed with publishers in the *Appendix*, pages 585-595.

TEST II (SEVENTEEN YEAR STANDARDS)

	Boys	Girls
1. Standing Broad Jump . .	6 feet 6 inches	5 feet 2 inches
2. Ball Throw (overhead)	39 feet	30 feet
3. 75 yard Dash	10⅖ seconds	50 yd., 8 seconds
4. Ball Throw (free style)	60 feet	45 feet
5. Chinning	6 times	1 time

TEST III (NINETEEN YEAR STANDARDS)

	Boys	Girls
1. Standing Broad Jump . .	7 feet 2 inches	6 feet
2. Ball Throw (overhead)	42 feet	33 feet
3. 100 yard dash.	13 seconds	75 yd. 11 seconds
4. Baseball Throw ; 75 ft., boys ; free style for girls	2 times	50 feet
5. Chinning	8 times	2 times
6. Swimming, free style	220 yards	220 yards

Carefully prepared rules are provided to assist in scoring each of the events. Awards are given on the attainment of the standards set up for each test.

At the beginning of each term, pupils are assigned to teams for track and field work, and official records are kept for each participant. Personal hygiene records are also kept. A carefully constructed table has been prepared to assist in recording numerical judgments of each pupil's performances and achievements. A *Merit Grouping Plan* is also provided for, the aim being to provide an ideal that correlates all the factors—physical, mental, and social, operating in the schools. Directions for the use of the plan are provided.

The *Detroit Decathlon for Boys*, devised by the physical education division of the Detroit Public Schools, has for its purpose the selection of the best all-around athletes in the school system. An annual meet is held to determine these persons. In this meet boys are eligible to compete for three different medals—gold, silver, and bronze. Before a boy is eligible he must qualify as follows in the chin, dip and sit up exercises, earning 650 points in each event for the gold medal, 560 points in each event for the silver and 460

points for the bronze. The total number of points which must be earned in the ten events of the meet to qualify for the gold, silver, and bronze medals respectively are 6500, 5600, and 4600 points.

The Playground and Recreation Association of America designed the *Athletic Badge Tests* because they felt that the facilities for promoting athletics and sports in the rural communities were decidedly inadequate. These tests are characterized by the fact that very simple equipment is required. The tests are as follows:

Test for Boys

First Test

1.	Pull-up (chinning)	4 times or
	Rope-climb (using both hands and legs)	12 feet
2.	Standing broad jump	5 feet 9 inches
3.	60-yard dash	9 seconds or
	50-yard dash	8 seconds
4.	Baseball throw (accuracy)	3 strikes out of 6 throws at 40 feet or
	Baseball throw (distance)	130 feet

Second Test

1.	Pull-up (chinning)	6 times or
	Rope-climb (using both hands and legs)	10 feet
2.	Standing broad jump	6 feet, 6 inches or
	Running broad jump	12 feet
3.	60-yard dash	8 seconds or
	100-yard dash	13⅗ seconds
4.	Baseball throw (accuracy)	3 strikes out of 5 throws at 45 ft. or
	Baseball throw (distance)	195 feet

Third Test

1.	Pull-up (chinning)	9 times or
	Rope-climb (using hands only)	16 feet
2.	Running high jump	4 feet, 4 inches or
	Running broad jump	14 feet
3.	220-yard run	28 seconds or
	100-yard dash	12⅗ seconds
4.	Baseball throw (accuracy)	3 strikes out of 5 throws at 50 ft. or
	Baseball throw (distance)	220 feet or
	8 lb. shot put	28 feet

Test for Girls

First Test

1. Balancing (1 deep knee bend)	24 feet, 2 trials
2. Potato race	22 seconds or
All-up Indian club race	30 seconds or
50-yard dash	8 seconds
3. Basketball throw (distance)	35 feet or
12-inch indoor baseball throw (accuracy)	2 strikes out of 5 throws at 25 ft.
4. Volleyball serve	2 in 5 or
Tennis serve	3 in 6 or
Basketball goal throw (ten foot line)	2 in 5 or
12-inch indoor baseball throw and catch	3 errors allowed

Second Test

1. Balancing (book on head; 1 deep knee bend)	24 feet, 2 trials
2. Potato race	20 seconds or
All-up Indian club race	28 seconds or
Run and catch	19 seconds or
50-yard dash	7 3/5 seconds
3. Basketball throw (distance)	45 feet or
12-inch indoor baseball throw (accuracy)	3 strikes out of 6 throws at 30 ft.
4. Volleyball serve	3 in 6 or
Tennis serve	3 in 5 or
Basketball goal throw (12 foot line)	3 in 6 or
12-inch baseball throw and catch	2 errors allowed

Third Test

1. Balancing (book on head; 3 deep knee bends)	24 feet, 2 trials
2. Potato race	18 seconds or
Run and catch	17 seconds or
50-yard dash	7 1/5 seconds
3. Basketball throw (distance)	55 feet or
12-inch indoor baseball throw (accuracy)	3 strikes out of 5 throws at 36 ft.
4. Volleyball serve	3 in 5 or
Tennis serve	3 in 4 or
Basketball goal throw (15-foot line)	3 in 5 or
12-inch indoor baseball throw and catch	1 error allowed

The rules for administering these tests are those which apply to any standard event. The alternate tests make possible some degree of choice in case equipment may be

limited. The battery is designed to measure all-around efficiency rather than particular types of achievement.

IV. Preventive and Corrective Work in Health and Physical Education

266. Prevention and Correction in Health and Physical Education. Preservation and improvement of pupil-health are two of the first duties of every teacher and administrator. Good health is basic to effectiveness in all phases of school and adult life. Therefore, an appraisal of each child's health status is a prerequisite to the planning of any program of work, academic or physical. Individual abilities and individual needs determine the character of instruction, the types of activities, and the speed of progress in health and physical education as in every other field of teaching.

Several administrative procedures are necessary to discover the health status of each pupil and to insure continuous protection of his health. Some of these are:

1. Medical examinations by competent physician to detect disease and other defects, remediable or otherwise.
2. Nurse service or follow-up service to secure remedies for defects and to give emergency treatment.
3. Daily inspection of all children in school, for cleanliness and to prevent the spread of disease.
4. Regular inspections of the sanitary conditions and safety of the school building, grounds, and equipment.
5. Examinations of health knowledge and information.
6. Physical ability tests of a performance character.

Such examinations would reveal cases of malnutrition, of defective hearing, defective vision, defective teeth, of goitre, of spinal curvature, of flat feet, of rheumatism, of adenoids and diseased tonsils, of mental or emotional abnormalities or disturbances. All of these tend to retard growth and learning as well as threaten the physical well-being of the individual. Without the proper treatment pupils so afflicted are almost certain to make unsatisfactory progress in both the physical and academic phases of their school work. Timely treatment is effective in most cases and fully pays its way for the time, effort, and money expended.

Complete diagnosis in health and physical education is not possible without the assistance of physicians, nurses and psychologists. Schools without such local services usually have access to county or state agencies.

Health training and instruction to remedy deficiencies in health habits and in health information should be based upon a careful analysis of such examination results. The remedial work to be effective must have certain goals: (1) to establish health habits and standards, emphasizing acts rather than the mere possession of knowledge; (2) to instill a working knowledge of practical facts relating to cleanliness and sanitation, food, fresh air, exercise, rest, sleep, and the causes of preventable diseases and the means of preventing and checking them; (3) to build ideals regarding health, beauty, and service for self, school, and community that will result in better living; (4) to develop the individual's sense of his personal responsibility, not only for his own health, but for that of the community in which he resides, and to set for the community standards of health.

Instruction and training in physical education should meet corrective, developmental and recreational needs. In all schools there are some children, who, either through disease, mental or physical weakness, or careless habits of posture, are badly in need of special gymnastic exercises. The ordinary program of calisthenics, dancing, games, and plays, does not correct these defects. The need is for special exercises which may safely be given as part of the corrective program, only by the expert in physical education, under the guidance and supervision of an orthopedic surgeon. Much may be done, however, even without this expert supervision, by a well-trained instructor, in the way of correcting flat feet, weakened abdominal muscles, round shoulders, faulty head-posture, and the ordinary slouch of the careless adolescent. Frequently some special equipment may be required for this work but such equipment need not be expensive.

Formal gymnastics have special and definite values if not carried to excess. They are useful in bringing about specific types of physical development. These exercises must be carefully graded to the age, strength, and physical and mental ability of the pupil. The choice of gymnastics and

their use requires a high degree of training upon the part of the physical instructor.

Physical education to be complete must meet the recreative as well as the corrective and developmental needs of children. Play is fundamental in the life of the child, and is a much more important factor in the life of the adult than most people realize. For adults it is a means of preserving health, both mental and physical, is a safe and natural outlet for the use of leisure time, is a source of joy and happiness, and offers an excellent field for those social contacts which are so significant.

Discussion Exercises

1. What are the real elements of distinction between health education and physical education ?
2. Which field, health instruction or physical education, offers the greater possibilities for objective measuring instruments ?
3. Through the use of the weight-height-age tables on pages 504 and 505 work out your own weight deviation for the age nearest your own.
4. Why has the emphasis in physical education shifted in recent years from gymnastics to games and sports ?
5. Work out a constructive program in which a proper balance is effected between the emphasis given to health instruction, and physical education and that given in most high schools to the competitive aspects of interscholastic athletics.
6. How many of the Brace Motor Ability Tests can you pass ?
7. Enumerate the special advantages of the Athletic Badge Test.
8. What is actually revealed by the Motor Quotient as proposed by McCloy ?
9. List in your own words the goals of corrective and preventive work in health training and physical education.
10. What is the place of play in meeting the developmental and corrective needs of pupils ?

Selected References

Baldwin, Bird T., *The Physical Growth of Children.* University of Iowa Studies in Child Welfare, Vol. 1, No. 1, Iowa City, Iowa : University of Iowa, 1921.

Baldwin, Bird T., Wood, Thomas D., *Weight-Height-Age Tables*

in English Units. Iowa City, Iowa: Iowa Child Welfare Research Station, University of Iowa.

Bovard, John F., and Cozens, Frederick W., *Tests and Measurements in Physical Education.* Philadelphia: W. B. Saunders Company, 1931.

Brace, David K., *Measuring Motor Ability.* New York: A. S. Barnes and Company, 1927.

Dansdill, Theresa, *Health Training in the Schools.* New York: National Tuberculosis Association, 1928.

Franzen, Raymond, et al., "Health Education Tests," *School Health Research Monographs,* No. 1. New York: American Child Health Association, 1929.

Health Education, A Program for Public Schools and Teacher-Training Institutions. Report of the Joint Committee on Health Problems in Education, 1930. Thomas D. Wood, Chairman, 525 West 120th Street, New York.

Keene, Charles H., *The Physical Welfare of the School Child.* Boston: Houghton Mifflin Company, 1929.

Kerr, James, *The Fundamentals of School Health.* New York: The Macmillan Company, 1927.

McCloy, C. H., "The Measurement of General Motor Capacity and General Motor Ability." *Research Quarterly Supplement,* Vol. 5:46-61, 1934.

McCloy, C. H., "Tests and Measurements for the Administrative Program of Physical Education," *The Journal of Health and Physical Education,* Vol. 3:9-11; Sept., 1932.

Nash, J. B. (Editor), *Interpretations of Physical Education,* Vols. 1, 2, and 3. New York: A. S. Barnes and Company, 1931.

Odell, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

Rogers, F. R., *Educational Objectives of Physical Activity.* New York: A. S. Barnes and Company, 1930.

Rogers, F. R., *Fundamental Administrative Measures in Physical Education.* Newton, Mass.: The Pleiades Company, 1932.

Rogers, F. R., *Physical Capacity Tests in the Administration of Physical Education.* New York: Bureau of Publications, Teachers College, Columbia University, 1925.

Smith, H. L., Wright, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

Symonds, P. M., *Measurement in Secondary Education.* New York: The Macmillan Company, 1927.

Wilson, G. M., Hoke, K. J., *How to Measure,* Revised. New York: The Macmillan Company, 1928.

Wood, Thomas D., and Cassidy, Rosalind R., *The New Physical Education*. New York : The Macmillan Company, 1927.

Wood, Thomas D., and Brownell, Clifford L., *Source Book in Health and Physical Education*. New York : The Macmillan Company, 1925.

Wood, Thomas D., and Rowell, H. G., *Health Supervision and Medical Inspection of Schools*. Philadelphia : W. B. Saunders, 1927.

CHAPTER XXI

THE MANUAL AND INDUSTRIAL ARTS

This chapter has as its purpose the discussion of the following points involved in the teaching and measurement of the manual and industrial arts:

a. General significance of the industrial arts.
b. Objectives of industrial education.
c. Objectives in household economics.
d. Tests in industrial education.
e. Tests in household economics.
f. Aptitude, vocational and trade tests.

I. Social and Educational Significance of the Industrial Arts

267. Manual and Industrial Arts in the Modern Educational Program. Education, in addition to its other responsibilities, must make itself immediately useful to those engaged in the process of living in a very practical world. The term "industrial education" itself emphasizes the need for an education which is a preparation of *all* for successful living in an industrial society, as well as a specific preparation of *some* for the more specialized types of productive labor in this same society. Thus the industrial arts face a rather delicate problem in meeting these demands.

Formerly, vocational courses were designed to prepare the individual for a more effective entrance into and pursuit of a particular kind of occupation. In our modern educational program, courses of this type are not particularly important due to the high state of specialization of the industrial system. Factories and plants have become so mechanized and the type of work that each workman is called upon to do is so specialized that there is little practical need for the genuinely vocational course. Furthermore, most industrial es-

tablishments have found that much of what is taught the student in the vocational courses must be unlearned before he becomes an effective workman in the narrow tasks he is called upon to perform. Practically all industrial organizations employing any number of workers maintain their own training schools in which the workers are taught their specific tasks under factory conditions. But this does not mean that industrial education is not receiving any emphasis in the general program of education. It merely means that the emphasis has changed. Specialization in the vocations and in manufacturing processes has made it more and more important that the school shall build up a rich pre-vocational background rather than technical skill. This the manual and industrial arts courses of the junior and senior high schools are now attempting to do.

The manual arts courses have a place in the general program of education corresponding to their effectiveness in helping men to become socially efficient. Such courses contribute to social efficiency in different ways. To some individuals they may give vocational power through actually contributing to the individual's ability to earn a livelihood. To all they should give first-hand knowledge of the material accessories of life. Man's physical welfare and happiness depend to a large degree upon his ability to select or construct, understand and use modern conveniences. Moreover, the manual arts in their modern form do much to develop an appreciation of beauty in relation to material, form, color, tone, and texture. This is only an element in aesthetic enjoyment but may contribute largely to efficiency and productivity.

In the elementary grades, the individual pupil is best served by the manual arts courses which give him a broad foundation of first-hand experience rather than specific vocational or technical training. The purpose here is to give each pupil contact with a large variety of materials and processes. It is not so much to teach techniques as to stimulate and guide his natural constructive activity and to utilize the opportunities which naturally present themselves at this age for expression through concrete materials. In the junior high school grades it may be assumed that the child is now

ready both mentally and physically to form definite habits and to develop definite skills in the use of his hands. The end sought may be vocational or general, or both, but in any case the manual arts should be so correct in technique, should place such emphasis on processes that are fundamental, should be so closely in harmony with the current forms of industry that they will have distinctly vocational values so far as they go.

A general summary of the industrial education courses is made peculiarly difficult by the number and types of courses. Among those usually offered in the various junior and senior high schools are manual training, or woodworking and cabinet making ; carpentry ; shop-work, which includes wood and metal working ; plumbing ; home mechanics ; elementary sheet metal ; electricity and electric shop ; printing ; automobile mechanics ; machine shop practice ; and mechanical drawing.

Industrial arts for girls are variously named. Probably the general designative title for such courses should be *Practical Arts for Girls.* Under such a general heading may be included domestic science, household arts, cooking, sewing, a general industrial course, bookcraft, home hygiene and care of the sick, and home decoration.

268. Objectives of Industrial Arts Instruction. Industrial arts courses must contribute in a very definite sense to the general educational objectives of the secondary school. The aims of industrial arts courses may be briefly stated as follows :

I. To offer exploratory industrial activities to aid in revealing interests, aptitudes, and vocational possibilities
   a. Such exploratory values are noted in the following "self-finding" aims :
      1. To try out individual interests, inclinations, and abilities for industrial pursuits through typical experiences
      2. To make reliable studies of the conditions, demands, and opportunities in related occupations
      3. To provide for the industrial needs of pupils who would not remain for academic education alone
      4. To help pupils more wisely to choose future courses in secondary and higher education

II. To provide general manual arts experiences of common value to all pupils who elect such work
   - *a.* Such common values are found in the following "consumers'" or utilizers' aims:
     1. To develop "handyman" abilities through repair and construction work for home, shop, and office
     2. To assist in better choice and use of industrial products and service
     3. To gain sympathetic attitudes toward other workers and their work
     4. To appreciate economic production by first-hand experience in production work

III. To offer opportunity for beginning specialized preparation for entrance into chosen industrial fields
   - *a.* Such opportunity should provide follow-up guidance and limited training facilities as are included in the following aims:
     1. To assist in the final selection of a promising career
     2. To extend the try-out activity to meet the preparatory-vocational needs of pupils who must leave school with a minimum of preparation
     3. To provide supplementary training through a sequence of co-ordinated shop work and related school work in mathematics, science, drawing, et cetera
     4. To offer greater opportunities for commercial experiences through co-operation with outside commercial and industrial agencies
   - *b.* The more intensive industrial courses to be formulated to serve two purposes:
     1. To meet the needs of boys with mechanical ability who might be better able to profit by industrial work than by the regular courses of study in the school program
     2. To assist those who for various reasons are to remain only a short time in school and who need short, intensive courses of training before entering industry

Throughout all industrial arts education the solving of problems, the training of individuals to meet new and unexpected situations intelligently, or training in creative or constructive thinking should be emphasized.

II. Tests in Industrial Education and Practical Arts

269. General Uses of Industrial Arts Tests. The possible uses of tests in the industrial arts fields are many. One very important use is in the prediction of probable future success in specific types of vocational work. The applications of general intelligence tests and tests of special aptitudes to this problem have been discussed in Chapters XI and XII of this volume. Tests in industrial arts are also useful for evaluating the outcomes of instruction. A few very useful tests for this purpose have been developed. Much work in the refinement of measurement techniques remains to be done here as in most other fields of instruction. Naturally, progress in the development of more effective measuring instruments must follow the more exact definition of the unit skills and informational elements comprising the minimal essentials in industrial arts instruction.

270. Measurable Factors. The following factors, affecting or revealing the achievement in the industrial or practical arts, lend themselves more or less readily to objective measurement. Educational guidance and instruction in the manual and industrial arts should be made much more effective through the measurement of these factors and the wise use of the results in the classroom.

Measurable Factors in Industrial Arts

1. General ability as indicative of probable success in the manual arts.
2. Special abilities as indicative of success in manual arts.
3. Interests and attitudes — mechanical, vocational.
4. General information — vocational, mechanical, technical.
5. Degree of trade abilities possessed.
6. Knowledge of diversified vocational activities.
7. Skills in mechanical drawing.
8. Artistic judgment.
9. Essential knowledge in field of home mechanics.

In the Home Economics field the following factors affecting or revealing degrees of achievement are at present measurable. These factors vary from specific, definite, and highly specialized skills to broad phases of appreciation.

MEASURABLE FACTORS IN HOME ECONOMICS

1. Specific skills, such as skill in certain phases of sewing, cooking, decorating, drawing.
2. Knowledge of important phases of home economics, such as knowledge of foods or household activities.
3. General information, in the matter of foods, clothing, household activities.
4. Aesthetic appreciation.
5. Artistic judgment.
6. Essential knowledge in field of home-making.
7. Knowledge and ability in cooking.

271. **Mechanical Drawing and Visual Arts Tests.**[1] The ability to visualize and to project an accurate and tangible reproduction of the mental image is undoubtedly highly related to successful work in certain aspects of the industrial arts. Accordingly, manual and industrial arts teachers have been much interested in the objective measurement of these abilities. The *McAdory Art Test*, the *Lewerenz Test in Fundamental Abilities of Visual Art*, and the *Meier-Seashore Art Judgment Test* are three objective standardized tests which contribute definitely to this field. The *McAdory Art Test* should appeal particularly to industrial arts teachers since the themes of the illustrations used were selected mainly from common objects in everyday life having some aesthetic significance, as furniture and utensils, textiles and clothing, architecture, painting and other graphic arts. A discussion of these three tests appears in Chapter XIX of this volume. Brief comments on five mechanical drawing tests as typical of the field are given in this section.

The *Badger Mechanical Drawing Tests* are intended as a general checkup on the pupil's accumulated knowledge of mechanical drawing rather than as a measure of his drawing ability as revealed by his neatness, accuracy, lettering, etc. Test 1 covers the use of tools, linework, dimensioning, lettering; Test 2, projection; Test 3, pictorial drawing. The tests do not have time limits, but the directions suggest that the test papers be collected after all but two or three of the slowest pupils have finished. Evidence of the validity and reliability of the tests is not given in detail. There is no

[1] All tests mentioned in this chapter are listed, with publishers, in the *Appendix*, pages 585-595.

adequate manual, and the tests do not have norms or adequate suggestions for interpreting and using the results.

The *Castle Mechanical Drawing Test* is divided into five sub-tests. The first sub-test requires the pupil to identify similar parts of an object in top and side views by matching corresponding numbers and letters. Sub-test two deals with dimensions; three with geometric terms; four with pencil technique; five with inking. The working time for the test is 41 minutes. The first three parts are objective in scoring, but the last two depend to some extent on the teacher's subjective judgment, although the scorer is provided with letter rating scales and six points are mentioned for rating the drawing.

No definite statement is given concerning the validity of the test which is based on an analysis of instructional materials and a long teaching experience in mechanical drawing. The coefficient of reliability is not reported. The test is not standardized but should prove useful for measurement of drawing achievement in the same manner that a teacher-made objective test would be.

Fischer's Mechanical Drawing Tests are in two parts. Part I is a test of technical information necessary in drawing, and no instruments are needed other than a pencil. Part II is a performance test and requires the use of drawing instruments. Either part can be given in a 45-minute drawing period. Part I is composed of four sub-tests and Part II of three sub-tests. Both parts of the test should be given since it is desirable to test information and performance. The two parts of the test are not equivalent forms but are divisions of the same test. The test has considerable diagnostic value as it enables the teacher to see where the pupils have succeeded and where they have failed. The problems in the test are rated according to difficulty. The claims for validity are based on the analysis of textbooks, blueprints, courses of study and in addition on a survey of 100 schools to find out what was being taught, time being devoted to drawing, etc.

The coefficient of reliability determined by giving the same test twice to 150 sophomores in high school was .79. This is quite low for a standardized test. The norms of

accomplishment based on 2500 students are indicated on the basis of minutes of instruction.

The Nash-Van Duzee Industrial Arts Test, Test II, Mechanical Drawing is designed to measure objectively performance in drawing as well as information about mechanical drawing. The test, suitable for use in both the junior and senior high school, is divided into Part I and Part II and is available in two closely equivalent forms equalized on the basis of the results obtained from 500 mechanical drawing pupils in the ninth and tenth grades. There is a manual of directions, an objective scoring key and a class record sheet. The scoring key also includes a scale for the rating of lettering abilities. Part I of either form can be written in the ordinary classroom with a pencil, but Part II requires the use of mechanical drawing instruments.

The claims for validity are based on the analysis of textbooks, courses of study and reference books and on a rating of the analysis by several hundred persons interested in teaching mechanical drawing. The reliability of the test was found to be .87 (correlating Form I with Form II) with 203 fifth and sixth semester students in mechanical drawing.

Norms. Median and percentile norms of accomplishment by semesters and minutes of instruction based on 2500 cases are given for the junior high school and the first two years of high school.

The purpose of the *Wells-Laubach Knowledge Test of Mechanical Drawing* is to measure the informational side of mechanical drawing. It is intended for use in the high school grades. The time for giving the test is 30 minutes and for scoring, 10 minutes.

The *Wright Achievement Test in Mechanical Drawing* is intended for use in high school or in beginning courses in college. The purpose of this test is to measure achievement in mechanical drawing. The author suggests that the test should be of value to the teacher of mechanical drawing in the following ways:

1. In the assignment of school marks, by furnishing a more objective and reliable basis than that afforded by the judgment of the teacher alone.
2. In the determination of promotions and failures.

3. In college, it may be used to pick out those students who have had sufficient work in high school to do advanced work in college.
4. In comparison of results obtained by the teacher of one class with the accomplishment of similar classes in the same school or in other localities.

The reliability coefficient is .93. The required working time for the test is about 60 minutes.

272. Standardized Industrial Arts Tests. The *Nash-Van-Duzee Woodwork Test* 1, *Scale A* is designed to measure the junior and senior high school pupil's knowledge of processes, tools, materials and information used in woodworking. Five different types of questions are used in the test with appropriate directions for each type. The validity of the test depends on an analysis of teaching content as obtained from courses of study, textbooks, surveys, reference books and trade analyses.

The coefficient of reliability based on 200 cases was found to be .86, using the chance-half method and employing the Spearman-Brown formula for estimation of the reliability of the whole test. Norms based on 3000 cases are given on the basis of semesters and number of minutes of instruction from the first semester of the junior high school through the first two semesters of high school. The manual has suitable directions for administering and scoring the test with suggestions for use of results in teaching and as a partial basis for assigning shop marks. This is the first really satisfactory standardized test of achievement to appear in the industrial education field.

Scale B of this test, the companion test of Test I, Scale A, is for the purpose of measuring the pupil's skill in manipulating hand woodworking tools. The test is suitable for measuring manipulative achievement in junior and senior high school hand woodwork. Nash and VanDuzee state in the manual that "the test aims to measure the pupil's understanding of directions involving frequently used woodworking processes and procedures, the reading of a working drawing, the selection of proper tools to carry out the specified work and the ability to use the tools selected to do the required work."

Three tests, covering the basic knowledges in the fields of

elementary woodworking, machine-shop work and printing, are available under the titles, *Wells Knowledge Test in Woodworking*, *Wells Knowledge Test in Machine Shop*, and *Wells Knowledge Test in Printing*. The testing time for each test is thirty minutes. Standards are available for the high school grades in which these subjects are commonly offered.

The *Hunter Industrial Arts Tests* comprise 32 short objective tests of the pencil and paper type. Each test includes 25 objective questions on the particular subject measured. The following is a list of these objective tests according to the subjects or parts of subjects tested:

WOODWORK

W-1	Tools test
W-2	Fastenings test
W-3	Comprehension test
W-4	Trade names test
W-5	Reading test for rule or scale
W-6	True-false test
W-7	Completion test
W-8	Building parts test for carpentry
W-9	Board measure test
W-10	Multiple choice test for wood and lumber
W-11	Reading test for framing square
W-12	Objective test for wood finishing
W-13	Multiple-choice test for carpentry
W-14	Multiple-choice test for pattern makers

MECHANICAL DRAWING

MD-1	Reading test
MD-2	Missing line test
MD-3	Lettering test
MD-4	True-false test

MACHINE SHOP

MS-1	Tool test
MS-2	Comprehension test
MS-3	True-false test
MS-4	Micrometer reading test
MS-5	Multiple-choice test

ELECTRIC SHOP

E-1	Symbols test
E-2	Objective test

AUTOMOBILE MECHANICS

AM-1	Parts test
AM-2	Multiple-choice test

PRINTING

P-1	Completion test

RELATED SUBJECTS

G-1	Shop English
G-2	Shop mathematics
G-3	Shop arithmetic
G-4	Geometry

Woodwork, printing, machine shop, and mechanical drawing are measured by the *Wells-Laubach Industrial Arts Tests*. The tests are all of the pencil and paper type and with the exception of one in mechanical drawing are made up of 100 true-false statements. Twenty-five minutes is the working time for the woodwork and printing tests, twenty minutes for the machine shop and thirty minutes for mechanical drawing.

The content of the tests parallels in a general way teaching practice, but no scientific means of determining validity is reported. No statistical data are given for the tests, but tentative norms for each of the four tests are reported, based on the median accomplishment of about 1000 students for four semesters.

273. **Home Mechanics Test.** The chief purpose of the *Newkirk-Stoddard Home Mechanics Test* is to measure in an objective and analytical manner the essential knowledge that the pupils should acquire from a well organized course in Home Mechanics. The test is divided into two closely equivalent forms, A and B. Each form contains 36 jobs, comprising half the outstanding jobs in home mechanics. If it is desired to use the test for diagnostic or analytical purposes both forms should be given to each member of the class.

Four criteria were used to establish the validity of the test:

1. Surveys to check the jobs on the basis of social utility.
2. Surveys of the actual teaching content of 75 representative schools.
3. Analysis of courses of study and widely used commercial job sheets which were based on surveys.
4. Selection of jobs with procedures representative of a class of jobs rather than of a single job.

TABLE 64

RELIABILITY OF A SINGLE FORM OF NEWKIRK-STODDARD HOME MECHANICS TEST

(40 Minutes' Testing)

Grade	No. in Sample			Standard Deviation		P.E. Score	
		Job	Point	Job	Point	Job	Point
7	50	.49	.86	2.5	26	1.2	6.6
8	50	.64	.89	3.0	29	1.2	6.5
9	50	.59	.85	3.2	24	1.4	6.3
7–9	150	.54	.86	2.9	27	1.3	6.8

TABLE 65

NORMS FOR NEWKIRK-STODDARD HOME MECHANICS TEST

May Testing (N=396)

	Jobs		Points	
	Form A	Form B	Form A	Form B
Mean	5.2	5.6	74.7	74.6
Median	4.8	5.2	77.9	73.0
Upper Quartile	7.0	7.6	95.8	96.3
Lower Quartile	3.2	3.1	60.5	54.5
Standard Deviation	2.9	3.4	25.1	27.9

As a result of this study 72 of the high-ranking jobs in home mechanics were selected and stated in specially designed procedure-type exercises. Later these items were

divided into two closely equivalent forms of the test. The test is objective in its scoring.

In Tables 64 and 65 are given some statistical evidence concerning the test. Table 64 presents reliability data on a single form of the test. Table 65 gives May norms based on 396 cases from grades 7, 8, and 9.

274. **Tests in Household Arts.** The tests in the household arts deal with three major phases of instruction, (1) foods and their preparation, (2) textiles, and (3) household management.

Foods. The *Streeter-Trilling Food Preparation Test* is a brief test of twenty-five multiple-choice items requiring a seventeen minute testing period. It is intended to measure both judgment and informational aspects of the subject, but items of the latter type predominate. The content of the test was obtained from an analysis of textbooks and courses of study used in the junior-high school grades and from questionnaires sent to supervisors in the field. Norms are available for grades seven to nine inclusive.

The *King-Clark Foods Test* is a sixty-item test covering the knowledges and abilities essential to the body of subject-matter which should be included in any course in foods and cooking. Standards are provided for grades six to twelve. A thirty minute testing period is required.

One of the most comprehensive of the tests in this field is the *Illinois Food Test* prepared by the test committee of the Illinois Home Economics Association. It is an information test specifically designed to measure the pupil's judgment and general understanding of ways and means of preparing food. There are one hundred carefully selected questions arranged under fourteen heads which measure respectively: (1) function and composition; (2) digestion and absorption of foods; (3) meal selection; (4) food preparation; (5) principles employed in preparation; (6) leavening agents; (7) food sanitation; (8) food economics; (9) preservation and storage of foods; (10) meal service and etiquette; (11) fuels and cooking apparatus; (12) housewifery; (13) food adjuncts; and (14) source of more common foods.

The test may be used as a survey test, or used for diagnosis,

each unit being given upon the completion of the type of work it tests. The scoring is completely objective. A 35 minute working period is required for the administration of the test.

Textiles. The *Frear-Coxe Clothing Test* is designed to measure the knowledge of clothing possessed by high school girls who are taking or who have completed courses in elementary dressmaking. It is expected that the test can be used at the end of any semester up to four semesters. The test measures the following phases of clothing construction: (1) the use of the sewing machine; (2) hand sewing; (3) simple processes of construction; (4) use and alteration of commercial patterns; (5) dyeing, spot removal, and general laundering; (6) mending and renovation of clothing; (7) hygienic aspect of clothing; (8) principles of design and use of color; (9) good taste in dress for individual types; and (10) ability to solve problems in economic buying of clothing.

It is the intent of the test not only to furnish an instrument for general survey purposes but also to measure proficiency in certain related groups of items. There are five parts in the one test booklet: Part I, Fundamentals of Construction; Part II, Care and Repair of Clothing; Part III, Hygiene of Clothing; Part IV, Appropriateness of Clothing; and Part V, Economics of Clothing. Each of these five parts may be scored separately and interpreted separately, permitting the test to be used as a diagnostic instrument. The working time on the entire test is 60 minutes.

The materials of the *Stevenson-Trilling Patterns Test* are divided into five parts. This division assists the teacher in making an analysis of the necessary performances in the correct use of patterns. The class (or a pupil) may be unable to (1) name the parts of the pattern—which would preclude work on most of the other tests, (2) understand the pattern lines, (3) understand the notches, (4) alter the patterns, or (5) lay the patterns correctly on the material. The total score on the five tests gives a good indication of the success of the class or individual in making profitable use of patterns, and the scores on the tests separately give a diagnosis of the points which are in need of further drill.

The purpose of the *Murdock Sewing Scale* is to measure general ability in hand sewing. The scale consists of 15 parts, each a photograph of a sample of stitching. Three different views of each sample are presented: (1) full-sized sample which shows hemming, basting, running stitch, backstitch, overcasting, and combination stitch; (2) the reverse side of overcasting, basting, the backstitch, and the seam; and (3) the reverse side of the hem, basting, and the hemming, running and combination stitches.

The fifteen samples range in merit from zero ability to the most expert ability, the difference between samples being practically equal. The use of these samples of increasing excellence provides a challenge to pupils, and considerable accuracy may be obtained in the grading of sewing products. Grade norms are available from the sixth grade to the second year of college, and age norms from ten to eighteen years.

Household Management. The *Engle-Stenquist Home Economics Test* measures the broader phases of home economics knowledge. It tests for knowledge necessary for healthful and right living, as well as for a knowledge of the principles underlying certain skills in foods, clothing, and household management. Separate tests have been prepared in the following subjects: Foods and Cookery; Clothing and Textiles; and Household Management. Each of these tests is issued in two equivalent forms, A and B. The tests are designed for grade five and on through junior high school.

275. **Aptitude and Intelligence Tests in Industrial Education.** Individuals enter wage earning careers with a great variety of educational preparation, with varying degrees of general ability and the special abilities, all of which will to a large degree determine their success in a particular occupation. It is important therefore that the school shall be supplied with specific information concerning the degree of general education, the amount of general ability, and the nature of the special abilities demanded by the various wage-earning activities which the young worker may enter upon leaving school. In a similar way the school is concerned with determining whether or not the high school student possesses (1) an adequate body of occupational information,

(2) an adequate amount of general education, (3) an adequate degree of general intelligence, (4) the special intelligence, and (5) the specific skills needed to succeed in these same jobs. Educational records will reveal the educational achievement of pupils; exploratory courses will indicate their interests in specific occupational fields; intelligence tests should be used to reveal the amount of training which it is practical for a pupil to take; aptitude tests will show whether or not an individual possesses the fundamental qualities without which he would be handicapped in a certain occupation; and, vocational or trade tests will indicate whether or not a pupil possesses the specific skills required for proficiency in a particular occupation.

Both intelligence and special aptitude tests attempt to measure native capacity, but the former deals with its general aspects. By means of the former, it is possible to say that an individual has the intelligence required for success in any one of a considerable group of occupations. By means of the latter, we seek to determine to what extent he possesses the special qualities which make it possible for an individual to do successfully and without undue difficulty the skills which characterize one occupation as being different from others.

The specific problems of using intelligence tests and aptitude tests for general guidance purposes are discussed in Chapters XI and XII of this volume. Numerous tests of both types are described and evaluated in these chapters.

276. **Vocational and Trade Tests.** Intelligence tests and special aptitude tests, both designed to measure native capacity, are primarily prognostic in their function. Their purpose is to aid in the suitable choice of and preparation for specific fields of endeavor. No preparation for or experience in any occupation is assumed before the pupil is subjected to a vocational aptitude test. Vocational and trade tests differ from these tests in that they are designed to determine an individual's ability to perform the work of a specific occupation. Training for or experience in the occupation, or both, are assumed, and the test is designed to measure the results of these or to find out what stage of progress the individual has achieved. This type of test is

valuable as an aid in determining whether an individual should be recommended for a particular occupation requiring known standards of proficiency. This kind of performance test has the advantage of being typical of occupational situations.

Trade tests were widely used during the war[2] for the purpose of classifying army recruits as to services they could best render. These tests were designed (1) to test each recruit's *ability to do*, and (2) his capacity for preparing to do. These tests measured in each recruit the degree of trade ability already possessed as a result of knowledge of, or experience in a trade. These army tests assisted in dividing all recruits into four grades of workers, (1) novices who possessed practically no trade knowledge or skill, (2) apprentices who showed a fair degree of proficiency in the trade, (3) journeymen, who could be relied upon to perform trade jobs satisfactorily, and (4) expert workers who could be expected to meet any emergency in the field.

Four forms of tests were developed for each of 80 trades. One form, the oral trade test, consisted of a series of 15 to 20 carefully chosen questions concerning materials, tools, machines, and processes used in the trade, how these are used, and the principles underlying their use. These questions required very brief answers, and were so formulated that only a workman who had had extensive experience in the trade could make a high score on the test. The second form of the test, a picture trade test, consisted of an equally careful picture selection of tools, machines, and processes of the trade, which were shown one after another to the subject. When a picture was shown, a simple question was asked to which the subject was expected to give a prompt answer. The third form of the test was an actual performance test in which the candidate did certain typical jobs of the occupation and was graded carefully on each part of the work done. The fourth form was a written group test, in which the persons had to read and answer in writing standardized questions about the trade. The four forms of the test were developed because it was believed that some trade work-

[2] For a complete description of the scientific method in army personnel, see J. C. Chapman, *Trade Tests*. Henry Holt and Company, New York, 1921.

ers would do better on one type of test than on another.

Standardized trade tests have not been widely used since the World War. However, trade unions have developed a few tests for use in many phases of trade work. These tests are used in some vocational and trade schools. The development of standardized tests of occupational proficiency suitable for use in rating those who pursue any kind of vocational preparatory training in public schools seems entirely practical, even though little progress in this direction has yet been made. The reason for the lack of trade-test development in the public schools is two-fold: (1) vocational education, as viewed by most educators, is not aimed to develop skilled tradesmen; (2) the need of standardization and objectivity of measurement has not generally been regarded as a serious need.

Before tests can be developed to determine adequately occupational proficiency much work must be done to analyze and determine specifically those skills which should be the trade equipment of each worker. In this connection the reader is referred to three volumes on occupational analyses entitled *Opportunities and Requirements in Local Occupations*,[3] prepared and edited by A. N. Edgerton and others, for the guidance work in Detroit, Michigan. While this information applies to local industries, the analysis of trade requirements is sufficiently general to be of great value wherever the same trades are found. The methods of analysis are as significant as the particular analysis itself, for they suggest the similar techniques which may be followed by other workers engaged in making occupational analyses for their own local industries.

The demand for vocational education under school conditions has grown rapidly in the past two decades. There seems to be no reason for expecting this demand to diminish in the years immediately ahead. Rightly organized, vocational education will prove a profitable investment for society and will contribute to the making of the citizen as well as the worker. In the course of the development of a progressive social economy, we may expect it to be made obligatory upon every individual to acquire a certain amount

[3] Published by the Detroit Board of Education, 1922-23.

of vocational education. In such a movement there will be an increasing need for objective means for evaluating instructional content and for evaluating the results of its presentation. The construction, refinement and use of tests in this field will therefore be one of the most significant challenges to forward-looking workers in industrial education.

Discussion Exercises

1. In what specific ways are the demands made on industrial education today different from those of ten years ago? Why?
2. What in your opinion will be the effect of further mechanization of industry on the objectives of instruction in the industrial arts as stated in this chapter?
3. Expand the list of measurable factors in industrial arts or in home economics (page 526) depending upon your interests in the fields.
4. What are the arguments favoring the inclusion of the study of Visual Art Tests as important measures of industrial arts outcomes?
5. What do you consider to be the greatest weaknesses of the tests in Industrial Arts (or Home Economics as your interest may dictate)?
6. What particular merit do you see in the validation of testing content as done in the Newkirk-Stoddard Home Mechanics Test?
7. What future do you see for vocational aptitude tests? Defend your position. What are the differences between trade tests and vocational aptitude tests?
8. Why are local occupational surveys of great importance for curricular and guidance purposes?

Selected References

Badger, Alex J., *Standard Tests in the Fundamentals of Mechanical Drawing*, Tests 1, 2 and 3. Bloomington, Illinois: Public School Publishing Company, 1929.

Bennett, Charles A., *History of Manual and Industrial Education up to* 1870. Peoria, Illinois: Manual Arts Press, 1927.

Briggs, T. H., *Curriculum Problems.* New York: The Macmillan Company, 1927.

Castle, Drew W., *Mechanical Drawing Test.* Peoria, Illinois: Manual Arts Press, 1927.

Chapman, J. Crosby, *Trade Tests; The Scientific Measurement of Trade Proficiency.* New York: Henry Holt and Company, 1921.

Cleeton, Glen U., "Printing Tests for the Junior High School," *Industrial Arts and Vocational Education Magazine,* Vol. 19: 329. September, 1930.

Donson, George C., "A Machine Shop Test," *Industrial Arts and Vocational Education,* Vol. 20: 132-3, April, 1931.

Edgerton, A. H., *Vocational Guidance and Counseling.* New York: The Macmillan Company, 1926.

Fischer, Ferdinand A. P., *Mechanical Drawing Tests.* Milwaukee, Wisconsin: Bruce Publishing Company, 1930.

Flaherty, Edward B., "Electrical Shop Tests for Grades 7, 8 and 9," *Industrial Arts and Vocational Education,* Vol. 19: 142, April, 1930.

Flam, August, "Some Mechanical Drawing Tests," *Industrial Arts and Vocational Education,* Vol. 19: 150, April, 1930.

Hjertstedt, W. G., "Auto Mechanics Test," *Industrial Arts Magazine,* Vol. 18: 239-40, June, 1929.

Hull, C. L., *Aptitude Testing.* Yonkers, New York: World Book Company, 1928.

Hunter, Wm. L., *Shop Tests.* Peoria, Illinois: Manual Arts Press, 1927-1930.

Kitson, Harry D., *The Psychology of Vocational Adjustment.* Philadelphia: J. B. Lippincott Company, 1925.

Mays, Arthur B., *The Problem of Industrial Education.* New York: The Century Company, 1927.

Nash, Harry B., and Van Duzee, Roy R., *Industrial Arts Tests.* Milwaukee, Wisconsin: Bruce Publishing Company.

Newkirk, L. V., and Greene, H. A., *Tests and Measurement in Industrial Education.* New York: John Wiley and Sons, Inc., 1935.

Newkirk, L. V., and Stoddard, G. D., *Home Mechanics Test.* Iowa City, Iowa: Bureau of Educational Research, University of Iowa, 1928.

Odell, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

Smith, Homer J., *Industrial Education: Administration and Supervision.* New York: The Century Company, 1927.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

Snedden, David, and Warner, William E., et al., *Reconstruction of Industrial Arts Courses.* New York: Bureau of Publications, Teachers College, Columbia University, 1927.

Trade and Industrial Education. Bulletin 17, Federal Board for Vocational Education, Washington, D. C., 1924.

Weaver, C. G., "Trade Tests: Their Construction, Use and Possibilities in Industry," *Industrial Arts Magazine,* Vol. 10: 163-6, May, 1921.

Wells, G. K., *Shop Tests.* Peoria, Illinois: Manual Arts Press, 1929.

Wright, J. C., and Allen, C. R., *Administration of Vocational Education.* New York: John Wiley and Sons, Inc., 1926.

Wright, J. C., and Allen, C. R., *Supervision of Vocational Education.* New York: John Wiley and Sons, Inc., 1926.

CHAPTER XXII

MEASUREMENT OF GENERAL EDUCATIONAL ACHIEVEMENT

THIS chapter treats the following points in the measurement of general educational achievement at the high school level:

a. Advantages of general achievement batteries for survey purposes.
b. Limitations of general measures of achievement.
c. General vs. specific surveys.
d. Types of achievement batteries.
e. Some distinctive features of certain achievement batteries.

I. GENERAL MEASURES OF ACHIEVEMENT

THE emphasis throughout this volume is so definitely on the use of tests for analytical purposes that to some this chapter may appear inconsistent. However, a consideration of the practical problems of measurement in the high school classroom results in the conviction that there is a real service to be rendered by properly balanced general survey tests of achievement.

277. **General vs. Specific Diagnosis.** The battery-type of general achievement test affords a survey of the total instructional situation in the field for which it is designed. It presents a perspective of the measurable aspects of pupil accomplishment, and points out general areas of weakness and strength which need much more detailed and analytical study. A general survey test of any of the types described in this chapter may reveal a weakness in the language skills, for example. To the critical teacher this is a challenge to discover more exactly the factors underlying this limited achievement. Accordingly, the few cases identified by the general achievement test are subjected to a detailed analytical

test in the subject for the purpose of locating specific difficulties and their causes.

It is important that even in the high school field the teacher shall keep clearly in mind the fact that measurement of general achievement is not a substitute for more analytical types of measurement. It is merely the stepping-stone to it. It cannot be repeated too emphatically that throughout all of the supervisory and instructional uses of tests, measurement which is vague and indefinite is likely to be useless. In general, measurement which does not aid in the identification of individual pupil difficulties and their causes is futile, and any consistent correction of educational deficiencies without the assistance of exact analytical and diagnostic measures verges on the miraculous.

278. **Advantages and Disadvantages of the Battery-Type Tests of General Achievement.** Among the specific qualities of the battery-type tests of general achievement are the following which have been given considerable emphasis by persons interested in the improvement of classroom measurement.

Curricular Content. For the junior-high-school and upper-elementary grades the general achievement battery appears to afford a very satisfactory basis for sampling curricular content. In the high school grades, on the other hand, with their wider range of subject matter and electives, the special subject test may open possibilities for measurement in the subject which no survey test can afford. The practice of subject election in the high school creates a problem in the selection of content which is yet to be solved by the makers of general achievement tests. One plan, that followed by the authors of the *Sones-Harry High School Achievement Tests* and the *Iowa High School Content Examinations*, proposes to limit the subject-matter sampling to four major high school fields likely to be encountered by every high school pupil sometime during his four years of work. For the *Iowa High School Content Examination*, which is especially designed for senior survey and college freshmen testing, this limitation of the fields and the requirement that each student respond to the exercises in each field without respect to the contact he may have had with

the subject in his high school course, is not a serious problem. In the case of the *Sones-Harry High School Achievement Test*, which is intended to be used in all four years of the high school course, the matter assumes more serious aspects. Perhaps the solution of the problem lies in the development of a high-school survey test of multiple parts sampling into five or six different subject-fields and requiring the pupil to select a specified number of test parts, the ones selected depending upon the courses he has elected. This procedure would more adequately meet the problem of the wide range of subject selection encountered in most high schools. A survey battery built along these lines would certainly do much to strengthen the curricular content of general achievement tests in the high school field.

Units of Measurement. The use of a uniform unit of measurement in the scaling of battery tests constitutes a real advantage in the interpretation of the test results and in the comparisons of results from one subject-matter field to another. While this is an important advantage it does not at all mean that uniformity in units of measurement may not be secured in single tests in unrelated subjects. As a matter of fact special scales for the commensuration of the scores from the various parts of battery tests are in common use. Examples of test batteries using these devices are seen in the *New Stanford Achievement Tests* and in the *Unit Scales of Attainment.*

Ease of Administration. The tendency of the authors of battery-type tests to utilize the same or similar types of testing techniques throughout the series of tests unquestionably tends to simplify the problems of administering the test. The use of uniform methods of recording the pupil's responses also simplifies the pupil's problem, and definitely reduces the difficulty of scoring the tests. In general, however, the length of such battery tests is so great that the problem of scoring the tests becomes quite serious. One of the very significant criticisms of the *New Stanford Achievement Tests*, the *Metropolitan Achievement Examinations*, the *Unit Scales of Attainment*[1] or any of the other

[1] A complete list of the publishers of all tests mentioned in this chapter will be found on pages 585 to 595 of the *Appendix.*

longer batteries suitable for junior high school use is the fact that the time required for administration and for scoring is extremely long. It will be noted that the *New Stanford Advanced Examination* requires four sittings of 42, 50, 45, and 33 minutes each. Some experience in the use of this and similar batteries of general accomplishment indicates that it requires a speedy clerk to score more than four of these advanced tests in an hour. However, it may be that this is not too high a price to pay for extensive sampling and reliable measurement.

Simplicity of Interpretation. The use of comparable units of measurement and similar testing techniques in the several tests comprising a general achievement battery simplifies the problems of comparing and interpreting the results. The raw test scores are readily turned into scale values, educational ages, grade equivalents, etc. Modern graphic methods of summarizing test results make effective use of such derived scores. Profile charts of the type used with the *Metropolitan Achievement*, the *New Stanford Achievement*, the *Unit Scales of Attainment*, the *Progressive School Achievement Tests*, and the *Modern School Achievement Tests* add to the clearness with which such test results may be interpreted.

Unity of Population in Standardization. The fact that the standardization of most comprehensive batteries is based upon returns from the same individual pupils for each of the different subject-tests insures a better picture of the relationships of achievement in these different subjects. The relation of achievement in reading for pupils of a given age or grade to the language achievement of pupils of the same age and grade can be obtained only when tests are standardized under these conditions.

Economy. Any economy in testing which results from the use of battery tests appears to be conditioned by the assumption that measurement is preferably broad and general in its scope and never specific. It is probably true that almost any one of the modern batteries of achievement tests will furnish a wider sampling into more subject-matter fields at a lower per capita cost than could be accomplished by the selection of single subject tests for the purpose. However,

there are numerous occasions, especially in the high school, when it is of greater importance to measure more intensively a limited range of subjects. For this type of measurement general attainment tests are clearly not the most economical. In order to provide for this situation the authors of most test batteries have prepared the tests for certain subjects in separate form. The main difficulty here lies in the fact that the use of any adequate number of these special separate tests runs up the total cost of measurement more rapidly than would result from the careful selection of more comprehensive tests of single subject-matter fields. Furthermore, most such tests are limited to use in or below the junior high school. After all, economy in measurement must be considered in terms of the net cost per unit of valid and reliable information secured. For some purposes the battery test may be most economical and for other purposes the choice will lie in another direction.

II. Types of General Achievement Batteries

A description of general achievement batteries written five years ago would, of necessity, have confined its attention to one or two such tests. Now there are many general batteries, several of which have very distinct merit. Only a few of the better known and more widely used high school achievement batteries are described here as types.

279. **The New Stanford Achievement Tests.** The original battery of *Stanford Achievement Tests* designed by Kelley, Ruch, and Terman and copyrighted in 1923, represents one of the outstanding measuring instruments of that period. These tests set new standards of validity, reliability, sampling, testing technique, standardization and interpretation for later workers. They undoubtedly did much to stimulate the improvement of measurement in general. After six years, and on the basis of much critical analysis, observation, and experimentation, these tests were revised in the form now known as the *New Stanford Achievement Tests.*

In their present form the *New Stanford Achievement Tests* are available in a primary examination for use in grades two and three, and in an advanced examination for grades

four to nine inclusive. Each examination is available in five forms, V, W, X, Y, and Z. Table 66 gives the names of the tests, the number of items in each test, the testing technique used, and the working time limits of the ten tests comprising the advanced examination which may be used in grades 4 to 9 inclusive. The reliabilities for the total score on the advanced examination range from a low of .89 in grade 4 to a high of .96 in grade 8.

TABLE 66

NEW STANFORD ACHIEVEMENT TEST: ADVANCED EXAMINATION

Test No.	Test Name	Number Items	Testing Technique	Working Limit
1	Paragraph Meaning	80	Compl.	25
2	Word Meaning	80	MC-5	10
3	Dictation Exercises	108[1]	Dictation	15[2]
4	Language Usage	74	Alternate	10
5	Literature	80	MC-3	10
6	History and Civics	80	MC-3	10[3]
7	Geography	80	MC-3	10
8	Physiology and Hygiene	80	MC-3	10
9	Arithmetic Reasoning	40	Problems	20
10	Arithmetic Computation	60	Examples	30[4]

[1] Total score for 9th grade
[2] Begins second sitting
[3] Begins third sitting
[4] Begins fourth sitting

280. The Unit Scales of Attainment. The *Unit Scales of Attainment* are a relatively new development in the measurement of general achievement. While the general function of these tests is practically identical with that of the *New Stanford Achievement Tests*, the *Metropolitan Achievement Tests*, etc., the actual methods used in the scaling and arrangement of the items are somewhat different. The subjects measured, the number of items in each test, and the testing techniques used are shown in Table 67. Nine different subjects are sampled with from twenty-five to forty

items per test. The multiple-choice type of exercise predominates, although the proof-reading or recognition-correction form is used in the English tests.

TABLE 67
UNIT SCALES OF ATTAINMENT

Test Name	Number of Items	Testing Technique
Reading	40	MC-5
Geography	40	MC-5
Literature	40	MC-5
Elementary Science	40	MC-5
American History	40	MC-5
Arithmetic Problems	25	Problems
Arithmetic—Fundamental Operations	25	Examples
Spelling	40	Dictation
English—Capitalization	30	Recognition
English—Punctuation	30	Proof-Reading
English—Usage	30	Proof-Reading

One of the distinctive features of the *Unit Scales of Attainment* is the arrangement of the scaled items in the divisions of the test. The tests, which are designed for use in grades three to eight inclusive, are arranged in three divisions, Division 1 for the third and fourth grades, Division 2 for the fifth and sixth grades, Division 3 for the seventh and eighth grades. There is no primary or advanced examination, but each division contains items of difficulty suited to the ability of the grades in which it is to be used. This is shown much more clearly by the accompanying analysis of the reading test items (Table 68). Division 1 contains ten reading comprehension paragraphs representing forty comprehension exercises. Paragraphs 5 to 10 inclusive are used in Division 2 as Paragraphs 2 to 7. These same paragraphs also appear as Paragraphs 1 to 6 in Division 3. Easier items are used at the beginning of Division 1 to afford simple enough stimuli for the lower grade pupils. Paragraphs 7 and 8 of Division 3 are also especially difficult

in order to afford adequate top for the advanced grades. The net effect of this scaling process and shifting of items into earlier stages of the test for the advanced grades is to produce a test which contains a larger amount of test material suitable for use in a specific grade than is possible otherwise. This plan of shifting test items is not followed exactly in all of the tests in this battery but the principle holds in most parts.

TABLE 68

ILLUSTRATION OF ARRANGEMENT OF READING TEST ITEMS UNIT SCALES OF ATTAINMENT

Reading Paragraph	Division 1 Grades 3-4	Division 2 Grades 5-6	Division 3 Grades 7-8
1	1		(1)
2	2	(2)	(2)
3	3	(3)	(3)
4	4	(4)	(4)
5	(5)	(5)	(5)
6	(6)	(6)	(6)
7	(7)	(7)	7
8	(8)	(8)	8
9	(9)	(9)	
10	(10)		

Each of the eleven tests in this battery is standardized on a mental-age basis. Raw test scores in all cases are changed into C-scores[2] by means of specially prepared scales. The individual profile charts simplify the graphic comparison and interpretation of the test results.

281. Metropolitan Achievement Tests. The *Metropolitan Achievement Tests* represent a very recent and ambitious development in the measurement of general educational accomplishment from the first grade through the eighth. The

[2] See discussion of C-scores on page 238 of this book.

series of tests is divided into four parts, each part covering no more than three grades. This insures more adequate samplings of material for each grade. The reliability of each single test as well as the entire battery is reported as being exceptionally high. Five alternate forms of the tests for the upper grades are available.

282. **Modern School Achievement Tests.** The *Modern School Achievement Tests* are similar in most respects to the other batteries previously discussed in this chapter. Table 69 gives a summary of the fields of subject matter measured by these tests, as well as the reliability coefficients reported by the authors for the upper grades.

TABLE 69

RELIABILITIES OF THE MODERN SCHOOL ACHIEVEMENT TESTS, GRADES 6, 7 and 8

Test	6	7	8
Reading Comprehension	.90	.90	.83
Arithmetic Computation	.87	.87	.84
Arithmetic Reasoning	.91	.90	.83
Spelling	.93	.96	.94
Health Knowledge	.95	.90	.91
Language Usage	.83	.79	.82
History and Civics	.79	.87	.93
Geography	.81	.86	.86
Science	.80	.76	.83
Entire Battery	.96	.97	.95

283. **Progressive School Achievement Tests.** This test is built in four divisions covering the primary, elementary, intermediate grades and the high school and college levels. The advanced battery is standardized for grades 9 to 13 inclusive. It provides an intensive sampling of 420 items covering reading vocabulary, reading comprehension, mathematical reasoning, mathematical fundamentals, and language. Data on the test are given in Table 70.

284. **Sones-Harry High School Achievement Test.** This test is unique in the fact that it was the first general achievement battery standardized for use wholly in the high school

TABLE 70

DATA ON PROGRESSIVE SCHOOL ACHIEVEMENT TESTS

Test	Number of Parts	Number of Items	Reliabilities (lowest; one-grade range)
Reading Vocabulary	4	100	.90
Reading Comprehension	3	55	.83
Mathematical Reasoning	4	60	.86
Mathematical Fundamentals	4	80	.84
Language	6	125	.89
Total	21	420	.96

grades. Norms are given for grades nine to twelve by semesters, and also by test parts in terms of the number of semesters the subject has been studied by the pupil. Data on the general organization of the test content and reliability coefficients are given in Table 71.

TABLE 71

DATA ON SONES-HARRY HIGH SCHOOL ACHIEVEMENT TEST

Part	Subject Tested	Number of Items	Relia-bilities
I	Language and Literature; word meaning, usage, authorship, grammatical and rhetorical forms, etc.	140	.94
II	Mathematics; processes, concepts, relationships, formulae, etc.	80	.91
III	Natural Science; principles, classification, scientific men and women, etc.	80	.90
IV	Social Studies; civic information, historical characters, international affairs, etc.	115	.93

285. **Iowa High School Content Examination.** While not specifically designed for high school use this test has had a wide demand for senior survey and college freshmen testing

purposes. The items appearing in the four sections of each form are sufficiently general in their scope to meet high school curriculum conditions in practically all parts of the country. All items are written in five-answer multiple choice form. The reliabilities of Forms A, B, A-1, and B-1 based on results from a large sampling of high school seniors are .95, .94, .96, and .95 respectively. Forms A-1 and B-1 are abbreviated arrangements of Forms A and B requiring a total testing time of fifty-five minutes as compared with eighty minutes for the longer forms. For high school testing purposes the shorter forms are probably more desirable.

Table 72 summarizes a few of the more useful items of information regarding these tests.

TABLE 72
DATA ON IOWA HIGH SCHOOL CONTENT EXAMINATION

Part	Subject	Forms A and B		Forms A-1 and B-1	
		Number of Items	Testing Time (min.)	Number of Items	Testing Time (min.)
1	English	110	20	75	15
2	Mathematics	75	20	50	15
3	Science	100	20	50	10
4	Social Science	115	20	75	15
	Total	400	80	250	55

286. **Iowa Every-Pupil Tests.** For a number of years special subject tests sampling into practically every phase of the high school curriculum have been given to from 50,000 to 75,000 pupils in the state of Iowa. While these tests are not designed to displace general survey batteries, their use throughout a school system gives a most illuminating picture of the general status of instruction and achievement. A few items of general information regarding the 1935 series of these tests are summarized in Table 73.

TABLE 73

DATA ON IOWA EVERY-PUPIL TESTS

(1935 Series)

Test and Subject	Number of Parts	Testing Time in Minutes	Number of Items
Algebra	3	60	56
American Government	2	60	101
Biology	1	60	112
Contemporary Affairs	1	60	105
Economics	2	60	126
English Correctness	1	55	180
General Science	1	60	96
History, United States	3	60	105
History, World	3	60	110
Latin	4	60	179
Physics	2	60	91
Plane Geometry	3	60	50
Reading Comprehension and Literature	1	60	133

287. **American Council Co-operative Tests.** This series of tests is the result of a subvention by the General Education Board. Their development has been under the general direction of Dr. Ben D. Wood. The collective efforts of a large number of scholars in special fields have been utilized in the preparation of this comprehensive series of tests. The program calls for the preparation of as many as ten or more comparable forms of examinations in the fundamental subjects at the senior high school and junior college levels. Thus far tests in English, literature, foreign language, mathematics, science, social studies, professional education and general culture have been announced.

288. **Other General Achievement Batteries.** Not all of the batteries of general achievement tests can be discussed here. Those described in this chapter serve merely as illustrations of tests of this type. Among the other general achievement tests which have rendered a real service in the intermediate grades in the past are the *Public School Achievement Tests*, the *Illinois Examinations*, the *Otis*

Classification Test, *Pintner's Educational Achievement Tests*, and *Lippincott-Chapman Classroom Products Survey Tests*.

Discussion Exercises

1. In your judgment what are the major supervisory functions rendered by the general achievement batteries in the high school?
2. Catalogue the specific advantages presented for batteries such as the *New Stanford Achievement Tests*, the *Unit Scales of Attainment*, or the *Metropolitan Achievement Tests*.
3. List in a similar way the objections to or disadvantages of such test batteries.
4. What is your own attitude with respect to the relative usefulness of such tests of general achievement as compared with a specific survey of a single high school subject?
5. How does the general survey battery fit into the diagnostic and remedial program?
6. Comment on the practical advantages of the arrangement of the items in the *Unit Scales of Attainment* as contrasted with the plan followed in the *New Stanford Achievement Tests*.
7. Who is likely to profit most from the use of general achievement tests, the classroom teacher or the supervisor?
8. Specifically, how may general achievement tests be used to objectify the grade placement and sectioning of pupils in a school system?
9. What is the advantage of the use of the C-score or any other standard score as in the case of a number of tests described in this chapter?
10. Secure from your instructor complete sample sets of the *New Stanford Achievement Tests*, *Advanced Examination*, the *Iowa High School Content Examination*, and the *Unit Scales of Attainment*, and after making a critical examination of each test, make a recommendation of one such test battery for use in a general survey in some high school grade, giving the reasons for your choice.

Selected References

Allen, Bixler, Conner, Graham, Hildreth, *Metropolitan Achievement Tests*. Yonkers, New York: World Book Company, 1933.

American Council on Education, *The Co-operative Tests.* New York : The Co-operative Test Service, 1933.

Gates, A. I., Mort, P. R., Symonds, P. M., et al., *The Modern School Achievement Tests.* New York : Teachers College Bureau of Publications, Columbia University, 1931.

Gilliland, A. R., Jordan, R. H., and Freeman, Frank S., *Educational Measurements and the Classroom Teacher*, Revised. New York : The Century Company, 1931.

Kelley, T. L., Ruch, G. M., Terman, L. M., *The New Stanford Achievement Tests.* Yonkers, New York : World Book Company, 1929.

Odell, C. W., *Educational Measurement in High School.* New York : The Century Company, 1930.

Tiegs, E. W., Clark, W. W., *Progressive Achievement Tests.* Los Angeles : Southern California School Book Depository, 1933.

Van Wagenen, Branom, Brueckner, Jordan, Cutright, Kelty, Anderson, Dvorak, *Unit Scales of Attainment.* Minneapolis : Educational Test Bureau, Inc., 1932.

Webb, L. W., Shotwell, Anna M., *Standard Tests in the Elementary School.* New York : Ray Long and Richard Smith, Inc., 1932.

Wilson, G. M., and Hoke, K. J., *How to Measure*, Revised. New York : The Macmillan Company, 1928.

CHAPTER XXIII

USING THE RESULTS OF TESTING

This chapter is the real objective of all of the preceding discussion of measurement in the high school classroom. It is concerned with three of the major problems involved in the testing program; namely,

1. Pupil adjustment through measurement.
2. Measurement of progress and improvement through the use of standardized tests.
3. Measurement of achievement by informal objective tests.

In general, this discussion assumes the point of view that any time or money spent in the construction or purchase of testing equipment is wasted unless the results revealed by the tests are returned to the classroom. Pupil- and teacher-time spent in any kind of testing situation is time wasted unless some tangible aid comes back to the classroom which will improve the conditions under which the teacher teaches and the pupils learn.

I. Pupil Adjustment Through Measurement

289. **Education as Guidance.** One of the most important outcomes of measurement in the secondary school is the establishment of a more adequate basis for the guidance of the pupil both in school and in his later life activities. The guidance function of education assumes a prominent place in the more modern high school. The teacher is the guide who accompanies the youth through his educational journey, and who is responsible for giving assistance when it is needed. It is assumed that the guide has been over the route before and knows the most important experiences, the most interesting and enlightening side-trips and the most economical short-cuts. Possibly it is too much to expect that many teachers will be able to serve as efficient

guides for pupils when they are often so inexperienced and poorly trained in the profession, so unacquainted in the ways of life that they themselves need guidance.

Inexperienced and sometimes poorly trained teachers and inadequate courses of study combine to make it important that reasonably reliable and valid supplementary devices be used to supply the essential guidance information. Ideally, the best possible guide a pupil could have is a wise, sympathetic, cultured, widely-travelled teacher with a sound philosophical background and a workable psychology of learning. The absence of such an individual in a school system makes it necessary to seek the required guidance in other quarters. Even the best judgment of the most sympathetic and learned teachers is subjective and is frequently wrong. Accuracy, objectivity, and validity in the data upon which the pupil's guidance is to be based are essential. Accordingly, it appears that the results of objective measurements have an important place in the educational program.

In a general way the guidance functions of education are represented in two forms, (1) the directive and (2) the corrective. The directive function of the school program is to provide the proper opportunities for the learning of the important activities which life is likely to call upon the individual to perform. The corrective aspects of education involve the more or less definite admission that something is wrong in the previous educational experience of the individual and that this difficulty can only be remedied through a further modification of the educational treatment he receives. This admission may be based upon an unfortunate assumption, but it is true that very few pupils pass through our school systems without at some time revealing a very definite need for corrective treatment of one form or another. The use of educational test data as the basis for the administration of the corrective program in education constitutes the major consideration of this section of this chapter.

Learning is fundamentally an individual and personal matter. Each person learns for himself through his own efforts, and each person probably learns most effectively in his own

peculiar way. It is the responsibility of the school and of the teacher to discover as far as possible these individual peculiarities and limitations of students. This responsibility is apt to be more keenly felt when the results of the students' efforts are unsatisfactory. When learning is accomplished with little friction and pain, little attention is given to it. Far more attention is given to the analysis and diagnosis of difficulties and causes of pupil weaknesses than to the more positive causes of success.

290. **Disciplinary Uses of Tests.** The experienced teacher, accustomed to handling pupil adjustments in the classroom, realizes that most teacher-pupil misunderstandings arise from inadequate information on the part of one or the other of the individuals involved. Probably by far the larger proportion of the problems of conduct in the high school classroom arise through the failure of the teacher to realize the actual abilities of the pupil and to utilize them adequately. The active, ambitious boy of superior capacity is almost certain to get into trouble unless his energy is directed into constructive channels. Too often the instructional material to which he is exposed is so uninteresting as to bore him. Many times he already knows as much about the subject as his teacher. In such cases it is small wonder that he gets into disciplinary difficulties.

Educational tests are not necessarily proposed as the cure-all for every type of disciplinary case, but it is quite certain the teacher through their use will find much help in securing a better understanding of such cases. Difficulties arising out of superior capacity inadequately utilized or stimulated may be anticipated through the analysis of mental test and achievement test results. Undue dissipation of time and energy through giving too much attention to extra-school activities may be revealed in terms of achievement results. Failure to achieve due to limited capacity may also be accounted for by such methods. Classroom teachers in all subject-matter fields through the use of properly selected educational tests may not only improve their understanding of the pupils, but they may also secure a most effective basis for the stimulation of the individual pupils to accomplish at their own most effective levels.

291. **Classification and Placement of Pupils.** Instructional efficiency within the classroom depends to a considerable degree upon the accuracy with which the material to be taught is adjusted to the ability of the learner. With the present rigid organization of curricular material in accordance with arbitrarily defined grade lines, it is quite important that pupils be classified within the grade in accordance with certain definite principles. Pupils will learn most effectively when placed with other pupils who have approximately the same initial abilities and who are able to learn at about the same rate. This means that pupils of approximately the same capacity and achievement levels should be grouped together for instructional purposes. For the accomplishment of this purpose the mental tests and general achievement tests naturally afford the basic data.

For classification purposes the use of the mental-age and achievement-age scores is recommended, since it results in placing together persons who have reached approximately the same levels of mental development. The ordinary basis of classification, the chronological age, disregards completely the actual power of the individual to learn. Grouping of pupils within narrow chronological-age limits actually results in a wide variation in mental ability. This is serious since it is the mental ability not the chronological age of the individual which principally determines his learning rate. Grouping pupils within narrow mental-age limits also results in wide variations in life age but this is not so serious. Other factors which should be considered in the classification and placement of pupils are the judgments of the teachers, and such objective evidence as may be gained concerning the physiological, social, and moral normality of the individual.

292. **Sectioning of Pupils.** A second factor in the placement of the pupil in the group for learning purposes is to place together individuals who are able to progress at approximately the same rate. While the mental age (plus data on the educational level of the pupil, and his physiological, moral, and social development) is very useful in locating the pupil as to general grade classification, it does not afford an adequate basis for sectioning the pupils within the grade. Every teacher knows that the instructional prob-

lems within the class are greatly increased when the range of ability is wide. Under these circumstances special efforts must be made to make the instruction simple enough for the least able student. To accomplish this for the slow student means that the more able ones are bored to distraction. Instruction adjusted, on the other hand, to stimulate the better students goes completely over the heads of the slow ones. Accordingly, it appears that in all cases of classes large enough in size to warrant more than one section for instructional purposes, the solution lies in placing together individuals who are at approximately the same level of relative mental maturity or are possessed of practically the same aptitudes in the subject.

Table 74

SCORES FROM A NINTH GRADE CLASS ON IOWA ALGEBRA APTITUDE TEST

Score		Pupils	
86 – 88	Fast Section	1	
83 – 85		1	
80 – 82		0	
77 – 79		3	
74 – 76		4	
71 – 73		3	
68 – 70		6	
65 – 67		6	
62 – 64		9	33
59 – 61	Average Section	10	
56 – 58		8	
53 – 55		10	
50 – 52		9	37
47 – 49	Slow Section	7	
44 – 46		9	
41 – 43		10	
38 – 40		6	
35 – 37		3	35
32 – 34	Divert from the course	0	
29 – 31		2	
26 – 28		1	
23 – 25		2	
20 – 22		0	
17 – 19		1	6
Total......		111	

An illustration of the manner in which results from a special aptitude test may be used for this purpose is given in the accompanying tabulation. The data in Table 74 represent the scores on the *Iowa Algebra Aptitude Test* from an entire entering ninth grade class in a small high school. According to the percentile norms for this test, pupils who score as low as the twentieth percentile are almost certain to fail or encounter great difficulty in algebra. The six individuals scoring thirty-one points or less may, therefore, very properly be diverted into other fields in which they have a more reasonable chance of successful accomplishment. This leaves 105 pupils to be divided into three sections. A practical division of these

pupils would be to place in a superior or more rapidly moving section the 33 pupils whose scores are 62 points or more on the aptitude test, and to place the 35 pupils scoring from 35 to 49 points in a more slowly moving section. The middle group of 37 pupils may be expected to make progress at about the normal speed for this subject. Such a plan should result in practically no pupil failures in the course, due to the elimination of the very poorly equipped individuals at the outset, and also to the very specific information relative to pupil weaknesses which the teacher will have in advance.

293. **Evaluation of Teaching Method.** Another very important use of educational test results lies in the evaluation of methods of instruction. As a matter of fact, it is only since the development of educational tests to their present state of refinement that any significant evaluation of teaching methods or materials has been feasible. Such a procedure naturally involves a rather carefully devised experimental situation in which all of the variables, with the exception of the one under observation, are controlled as far as possible. A preliminary measurement designed to equate the experimental groups or to reveal initial abilities must be taken. Following this the experimental factor is applied. A final measure of a type similar in every basic respect to the initial measurement is then taken. Differences in the initial and final results for the two groups under comparison are taken as indications of the operation of the experimental factor.

II. Measurement of Progress and Improvement by Standardized Tests

294. **Standards and Norms as Essentials in Measurement of Progress.** One of the features which distinguishes standardized tests from other objective measures of achievement is the fact that they are uniformly accompanied by norms or standards. Norms or standards, not only are tangible evidences of refinement in the development of the test, but they afford the basis for the practical interpretation of the meaning of the test results. The importance of the existence of suitable norms and the necessity for unusual care in their derivation and interpretation has been given

special emphasis in an earlier chapter in this book. The purpose here is to make it quite clear that they are essential in all cases in which the tests are to be used for the measurement of progress. Non-standardized tests and the subjective ratings assigned by teachers to essay-type examinations or even to objective classroom products provide no basis for the measurement of progress, since they have no point of beginning from which to express progress.

In subject-matter fields in which the units of instruction follow each other in rather systematic and logical order some work has been done recently in the development of narrow-unit tests which are standardized at the end of the period of instruction on the unit. Examples of tests of this type are the *Lane-Greene Unit Achievement Tests in Plane Geometry* and the *Iowa Unit Achievement Tests in First Year Algebra.* In certain other subjects which do not lend themselves to this type of analysis into narrow units, there is a growing tendency to provide norms which are adjusted to fit practically any period in the school course when the tests may be administered. Norms of this type make it possible to reveal progress and improvement in a very definite manner.

295. **Limitations of Standard Tests as Measure of Improvement.** The accuracy of measurement of improvement, growth or progress is essentially conditioned by all of the limitations inherent in the tests themselves. To these may be added certain limiting qualities which appear in the norms for the tests. No measurement can be more meaningful than the validity the instrument itself provides. Improper emphasis, inadequate samplings, ambiguous exercises, subjective scoring methods, improper working time limits, and doubtless many other factors limit the accuracy of all educational measurement. Measurement of improvement requires the use of test norms. Accordingly, any inherent inadequacies such as the lack of suitable groupings, a failure to select representative populations, or the use of insufficient population samplings, or errors in computation of the norms themselves, will also operate to limit the accuracy of measures of improvement as well as any other interpretations which may be attempted.

III. Measurement by Informal Objective Tests

296. Informal Objective Tests as Measures of Class Progress. The inadequacy of the standardized educational test to meet all types of measurement uses has been consistently emphasized in this discussion. There are certain functions which the standard test simply cannot perform effectively in the high school classroom. This limitation is inherent in its construction. The very fact that the standard test is designed to cover the common elements in the course of study makes it practically certain that the test will be unsuited for use in any class in which a different course of study is taught. Furthermore, the standard test is usually prepared by an individual other than the instructor of the class in which it is to be used. Thus, there is little or no assurance that the emphasis which the instructor has given to the subject matter will be reflected in the test content. Usually the standard test covers a much wider area of the subject matter than the instructor desires to cover in his periodical measurements. While these limitations of the standard test concern themselves mainly with matters of the validity of the test content, they are sufficient to invalidate its use for many purposes.

The problems of constructing and using informal objective tests have been given considerable attention in this volume. However, it may not be entirely useless to remind the high school teacher once more that without doubt by far the larger portion of time and effort he gives to the problems of measurement will involve the use of instruments other than standardized tests and scales. Whether or not the devices used are objective or subjective in character will largely determine the significance of the results.

297. Limitations of the Subjective Bases of Teachers' Marks. The case against the subjective evaluation of classroom accomplishment has been developed in current educational literature to the point that apparently little remains to be said. The evidence all points clearly to the fact that estimates based on personal impressions and unsupported by objective criteria may be expected to be quite unreliable. A

large part of this unreliability is accounted for by the absence of tangible units of measurement and of standards of expectancy on the part of the observer. The best evidence shows that a reasonably high validity may be found in many examinations utilizing the discussion-type question. The difficulty seems to be to secure an accurate estimate of the quality of the discussion. Something of the difficulty teachers have in the assignment of subjective marks to pupils' papers is demonstrated by the data in Table 75. This table summarizes the marks assigned by a group of 557 classroom teachers and teachers in training to a geography paper written by a sixth grade boy. This paper, in the opinion of the teacher of the class, was a passing paper; passing in this case being understood to be 75 or better. It is not the fact that the average of the marks assigned by these teachers is below passing which is the startling feature of this array. It is the extremely wide and consistent disagreement of the individuals as to the quality of the paper. Similar data selected from more exact fields such as arithmetic show distressingly similar results.

TABLE 75
MARKS ASSIGNED A 6TH GRADE GEOGRAPHY PAPER BY 557 INDIVIDUALS

Marks Assigned	No. of Teachers
89 – 91	1
86 – 88	4
83 – 85	5
80 – 82	17
77 – 79	14
74 – 76	40
71 – 73	57
68 – 70	95
65 – 67	102
62 – 64	75
59 – 61	66
56 – 58	40
53 – 55	24
50 – 52	9
47 – 49	4
44 – 46	2
41 – 43	1
38 – 40	1
Total	557

It is quite probable that the preparation and use of carefully evaluated answer keys and the administration of special periods of training in the grading of such examinations might eliminate some of the subjectivity, but it is a tedious and uncertain process.

Refinements in methods of evaluating pupil achievement which do not definitely place an objective unit of measure-

ment in the hands of teachers are makeshifts at best, and since it is apparent that by far the larger part of the evaluation of achievement is almost certain to be on the basis of teachers' judgments, it is important that really effective devices be used. For the time being at least it appears that the best solution of this problem involves the more extensive use of objective examination techniques and the wider popularization of simple statistical methods of turning such test scores into a meaningful teacher's mark.

298. **Functions of the Teacher's Mark.** Since the teacher's mark is now and almost certainly will continue to be the most frequently recorded measure of pupil accomplishment, it is very important that the classroom teacher have a definite notion of the functions of such marks. Teacher's marks function in a four-fold way: (1) they provide the basis for the school's record of the child's educational history, (2) they furnish the teacher with a record of the pupil's achievement and progress, (3) they reveal to the pupil the school's evaluation of his effort and accomplishment, and (4) they furnish to the parent reasonably accurate information concerning the pupil's achievement. For the school administrator, marks afford the common basis for determination of promotions, scholastic honors, and school classification. For the teacher, marks provide a working basis for group distinctions in assignments, work requirements, extra-curricular activities, etc. For the pupil, marks should give accurate information concerning the amount and quality of work done. For the parent, the marking system should supply accurate information on pupil achievement which should indicate relative success or failure in unmistakable terms. Obviously, the realization of these four-fold functions of the marking system places it under a very severe burden. The real severity of this burden is better appreciated when we recall the implications of the experimental evidence on the reliability of teachers' marks, and then in the face of these disturbing facts, realize the seriousness with which these marks are taken by the pupil, by the parents, and even by the school itself.

299. **Objectifying the Marking System.** A critical examination of the marking system and the marks assigned by

teachers makes it very apparent that some radical improvements in these phases of educational measurement are needed. As the result of an extensive survey of the problem, and a study of the recommendations of educators who have studied the problems of the marking system, the following program for eliminating many of the unsatisfactory features of the present methods of assigning marks is submitted. In brief, the plan is as follows:

1. *Discard the practice of marking pupils in percentages.* Three reasons are advanced for this decision: (a) The percentage scale has for its only fixed points 0 and 100. The former means just no ability while the latter means perfect mastery. Yet the complete scale is rarely used in practice. (b) The establishment of the limits of the scale fixes the intermediate values. Accordingly, the difference between marks of 75 and 76 should be the same as the difference between marks of 97 and 98. Common observation reveals the absurdity of this assumption. (c) The use of the percentage scale presupposes that the teacher is able to distinguish as many as 101 minute differences in accomplishment. Experimental evidence[1] reveals that teachers are able to distinguish from five to seven levels of ability. The use of a finer scale implies an exactness of discrimination on the part of teachers which does not exist. (d) The use of an arbitrarily selected percentage as a passing mark, as is very commonly done, results in throwing the marks into a badly skewed distribution with too large a proportion of the marks piled up at or near the passing mark.

2. *Each mark assigned to a pupil should be a symbol designed to indicate his power to do.* This symbol should be defined in exactly worded statements, understood alike by teachers, administrators, and pupils.

The following definition of letter grades given by Hillbrand (*School and Society* XXI, January 31, 1925, p. 142) is cited to illustrate the type of statements that may be used to describe the qualities of an individual deserving classification in each of the letter steps in the five-point scale.

[1] Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929; pages 370-374.

Grade	Definition
A	1. Consistently does more than is required.
	2. Has wide vocabulary at his command.
	3. Is always alert ; takes active part in discussions.
	4. Has unusual dependability in taking assignments.
	5. Is prompt, neat, and thorough in all work and unusually free from teachers' correction.
	6. Knows how to select books, tools, materials, and is a rapid worker.
	7. Has initiative and originality in attacking problems.
	8. Has ability to associate and re-think the problem and can adapt himself to new and changing situations.
	9. Has enthusiasm for and interest in his work.
	10. Has ability to apply ideas gained in study to everyday life.
B	1. Frequently does more than is required.
	2. Has good vocabulary and speaks with conviction.
	3. Unusually alive in the situation at hand.
	4. Careful in complying with assignment.
	5. Eager attack on new problems, profits from criticism.
	6. Prompt, neat, thorough, and usually accurate in all work.
	7. Has ability to apply general principles of the course.
C	1. Does what is required.
	2. Possesses a moderate vocabulary.
	3. Willing to apply himself during class hour.
	4. Does daily preparation with comparative freedom from carelessness.
	5. Attentive to assignments.
	6. Has ability and willingness to comply with instructions and a cheerful response to correction.
	7. Reasonably thorough and prompt in all work.
	8. Has average neatness and accuracy in all work.
	9. Has ability to retain collectively the general principles of the course.
D	1. Usually does what is required.
	2. Attendance often irregular.
	3. Tools and equipment sometimes lacking.
	4. Frequently "misunderstands" assignment.
	5. Willing but slow in complying with instructions and corrections.
	6. Careless in preparation of assignments.
	7. Lacking in thoroughness and sometimes tardy with work.
	8. Careless in presentation of work.
Fd	1. Usually does a little less than is required.
	2. Listless and inattentive in class.
	3. Tools and equipment for work often lacking.

4. Always tardy with work.
5. Seldom knows anything outside the lesson.
6. Retains only fragments of the general principles of the course.
7. Lacking in qualities of the first three groups to the extent that he cannot or will not do the work.

3. *Each teacher should give objective examinations or quizzes frequently throughout the term, and the scores from these tests should afford the major basis for her marks.* Prior to the assignment of marks for a school period or semester the pupils should be ranked on their test scores and these scores should then be transformed into marks on a five point letter scale by the use of the standard deviation technique in the case of large sections or classes (thirty or more). In the case of small classes (less than thirty) this may be accomplished somewhat more simply by dividing the distribution of scores into five groups and assigning the designated mark to previously determined percentages of the class. The letter grades used and the typical percentages of the class assigned each grade under these conditions are as follows:

Letter Grades	A	B	C	D	F_d
Percentage of Class	4–6	19–21	48–52	19–21	4–6

The essential steps in the assignment of grades by the standard deviation method are outlined in Chapter VIII. The actual solution of a problem utilizing this method in the assignment of marks to objective test scores from a class of forty-five pupils is shown in Table 18, page 184.

4. *Require teachers to prepare in advance for each six-weeks' period carefully worded statements of the objectives of each subject for that period.* Unless this is done, no one can determine whether or not the pupils are being tested over the things on which they should be tested. This statement of objectives should be the criterion by which the validity of the objective tests is determined.

5. *Work prepared for daily assignments should be treated as a requirement of the course, but marks assigned should be*

determined by numerous brief objective quizzes or tests over the work assigned.

6. *Notebook and laboratory work should be treated as a requirement of the course, and credit should be deducted or withheld for work which is unsatisfactory or incomplete.* However, the marks assigned should be determined by frequent objective tests over the work rather than on the basis of the notebook or laboratory work which may or may not be the pupil's own work.

TABLE 76

SUGGESTED POINT VALUES CORRESPONDING TO LETTER GRADES

Grade	Points
A+	16
A	15
A−	14
B+	12
B	11
B−	10
C+	8
C	7
C−	6
D+	4
D	3
D−	2
Fd	0

7. *Assign marks on "accomplishment" or "performance" rather than on indefinite subjective factors such as effort, attitude, ability, etc.*

8. *Final grades summarizing all of the quiz and test*

grades for the course can be obtained quite readily by assigning point values to each letter grade, computing the actual average for each pupil, and then re-assigning the final class marks on the basis of these averages. This is a very simple way of assigning final grades for fairly large groups and in courses in which a relatively large number of objective marks are to be summarized in the final grade. It also permits the weighting of certain period and final tests in accordance with the teacher's judgment of their importance.

Table 76 showing point values corresponding to specific letter grades may be useful to the teacher. Values are suggested for plus and minus values of the letter grades as one means of softening some of the shock from the arbitrariness of letter grades assigned on the basis of the normal curve. Students whose test scores fall just below the point where a superior grade is given sometimes feel that this is a distinct element of unfairness in the system. Assigning the plus and minus letter grades to their quiz scores serves to take care of this problem quite adequately.

Selected References

Bangs, C. W., and Greene, H. A., "Teachers' Marks and the Marking System," University of Iowa Extension Bulletin No. 244, May 15, 1930. Iowa City, Iowa.

Brueckner, Leo J., and Melby, E. O., *Diagnostic and Remedial Teaching.* Boston : Houghton Mifflin Company, 1931.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago : Scott, Foresman and Company, 1929.

Ruch, G. M., and Stoddard, G. D., *Tests and Measurements in High School Instruction.* Yonkers, New York : World Book Company, 1927.

Rugg, Harold O., "Teachers' Marks and Marking Systems," *Educational Administration and Supervision,* Vol. 1 : 117-142, February, 1925.

Odell, C. W., *Educational Measurement in High School.* New York : The Century Company, 1930.

Wilson, G. M., and Hoke, K. J., *How to Measure* (Revised). New York : The Macmillan Company, 1928.

Woody, Clifford, and Sangren, P. V., *Administration of the Testing Program.* Yonkers, New York : World Book Company, 1932.

CHAPTER XXIV

TESTS AND THE HIGH SCHOOL TEACHER

Summary

THE purpose of this chapter is to summarize the more outstanding problems involved in the use and interpretation of educational tests by the high school teacher, as presented in the foregoing chapters of this volume.

300. Recognition of the Need for Tests in the Classroom. Teachers long have measured the results of their teaching efforts. However, it is only relatively recently that any large degree of accuracy has been injected into their methods of measurement. For many years the teacher's estimate was accepted as the sole measure of a pupil's ability or accomplishment. Studies of the reliability of such methods gradually began to cast a doubt on their accuracy. Accordingly, interested teachers and research workers began to seek for more dependable measures. This movement was fostered by the so-called "survey" movement among educators which appeared at about the same time. Possibly the survey movement itself was a product of the same spirit of unrest and dissatisfaction with educational methods which brought into being the measurement movement. The survey movement left in its wake a perfectly logical result—the establishment of many centers of interest in the more exact evaluation of the results of educational practices. These later appeared as bureaus of educational measurements and research, a few of which are still functioning. Very distinctive service has been rendered by these agencies through their work in the construction, standardization, and critical evaluation of educational measuring instruments.

The testing movement has passed through the first stages

of its development. Fifteen years ago it was necessary to popularize the idea. Now the advantages of standardized and informal tests are recognized by most educators and by many laymen. Moreover, the tests themselves are being subjected by these same persons to careful analysis and refinement. The most enthusiastic students of educational tests are their own severest critics. The shortcomings of tests are coming to be realized. They are not mysterious instruments for the confusion of the uninitiated, but are useful devices for assisting the progressive educator to improve the conditions under which teachers teach and children learn. If they aid in the accomplishment of this they are justified.

301. **Meaning of Educational Tests.** The meaning of educational tests to the classroom teacher may be made clear most readily by considering the characteristics which distinguish them from other types of measuring instruments in education. In the first place, the educational test of standardized or semi-standardized form is more limited and specific in its purpose than the ordinary informal examination. That is to say, it usually confines itself to some relatively narrow field of subject matter while the typical examination or objective test prepared by the classroom teacher frequently covers a wide range of information. In the second place the exercises comprising standardized educational tests are commonly constructed and arranged in accordance with certain statistical and educational principles designed to produce more accurate measuring instruments. In the third place the more useful standardized educational tests have been used by a large sampling of school children under controlled conditions. From this use of the tests the norms and standards which give meaning to test results and permit the interpretation of test scores are derived. In the fourth place, the more carefully constructed and valuable educational tests yield results which point the teacher's way to the application of specific remedial methods where needed.

Educational tests are characterized by other features such as validity, reliability, objectivity, etc. Validity refers to the truth of the picture of the ability or achievement revealed by the test. Reliability refers to the consistency with which the test reveals this picture. Objectivity refers to the extent

to which the test results are affected by the personal judgment of the scorer. There are other important characteristics of educational tests but these represent the major ones on which their meaning depends.

302. Teacher-Made Measurements of Achievement. The emphasis on the use of the standardized test in most discussions of measurement problems often leads to the mistaken notion on the part of the student that these more formal types of tests are the most important measures of achievement. In most subject-matter fields this is distinctly not the case. The use of some form of testing procedure for instructional purposes probably constitutes nine-tenths of the teacher's measurement activity in the classroom. Accordingly, much more attention should be given to the improvement of the teacher's informal, traditional measures of achievement.

303. Using Informal Objective Examinations in the Classroom. The so-called new-type examination has increased in popularity with great rapidity in the past decade, although it is well recognized that there are certain aspects of educational accomplishment which it does not measure adequately. The advantages claimed for the informal objective examination are as follows:

1. The objective form of the stimulus used.
2. The brevity of the pupil's response.
3. The extent of the sampling which is made possible.
4. The resulting reliability of measurement.
5. The economy of pupil time in responding to the test and the teacher's time in scoring it.

Three of the more important criticisms of objective examinations which have been suggested are as follows:

1. They fail to give the student training in organized thinking and expression of his thought.
2. They tend to emphasize the measurement of factual items only.
3. They are expensive to prepare.

The answer to the first of these objections is that it has never been demonstrated that such training in organized thinking and expression is adequately provided in any other type of examination, and furthermore, it is too important an outcome to be left to such incidental influence as may come

from any examination situation. Some recent experimental evidence makes it quite clear that the emphasis on facts as such which results from the use of objective tests does not handicap the pupil in his general comprehension of principles.

The successful construction of objective examinations calls for the application of many of the same principles of test construction as are involved in the development of standardized tests. In many respects the informal objective and the standardized tests are quite similar in nature and in use. While results from non-standardized tests within certain limits yield information which may aid the teacher in making instructional adjustments within the class, the chief practical value lies in their use as the basis for class grades. Standardized test scores are not at all suitable for this purpose, but informal tests of the objective type afford the ideal basis for the objectification of the marking system.

304. **Uses of Standardized Educational Tests.** Educational tests, because of their definiteness and objectivity, reveal to the teacher the status of the achievement of his class. They point out individual differences in capacity and achievement. If standardized they set up specific goals of achievement for the teacher. They reveal the results of special types of emphasis, or of special methods of instruction. They open to the administrator and teacher hitherto untouched sources of information useful in giving the pupil proper educational and vocational guidance. In their modern conception they reveal to the teacher the specific weakness of individual pupils so definitely that the teacher is in a position to apply effective instructional and corrective methods. Tests themselves have little or no power to bring about changes in pupil achievement as a mere result of their use. Their chief service is their power to reveal pupil strengths or weaknesses. The correction of weaknesses is another aspect of the supervisory problem.

305. **Using Standardized Tests in the Classroom.** Teachers themselves must assume a larger share of the responsibility for the use of educational tests in the classroom, and for the interpretation and application of the results after the tests have been given. Only by so doing does the teacher receive an adequate return from the use of tests. If this

responsibility is to be wisely assumed the teacher must have an appreciation of the possibilities and the weaknesses of tests. He must be trained in their use and interpretation. He must be willing to exchange a certain amount of personal effort for the information concerning his teaching problems which the tests can furnish him.

Training in the use of tests comes as a result of their use. Opportunity for this training may be afforded through the preparation and use of objective informal tests as substitutes for the traditional examination, or it may be provided by undertaking a study of some supervisory problems of importance to the teacher in which standardized tests are used.

Main Uses of Tests. Three main types of uses of educational tests are noted, each resulting in different points of view regarding the teacher's responsibility. Tests of a detailed diagnostic type designed to give the teacher precise information concerning the abilities and limitations of his pupils are instructional in their function. The responsibility for the use of such material, in addition to informal objective tests for gradation purposes should be the teacher's. Tests designed to be used more particularly for survey or supervisory purposes should probably be administered by persons other than the teacher. The use of tests for administrative purposes such as for pupil classification, gradation, or sectioning, may well be the joint responsibility of the teacher and the administrator. In other words, the function the tests are intended to perform determines where the responsibility for their use and interpretation lies.

Selection of the Test to Use. The common criteria for tests, validity, reliability, objectivity, administrability, and economy, afford the teacher a tangible basis for the selection of the test to use for a certain purpose. In a general way, however, the teacher should depend upon the advice of persons who have made a special study of the tests, rather than attempt to apply these criteria personally. Information on each of these important aspects of tests is furnished with most of the better and more recently developed tests. If affirmative answers to each of the following questions are available concerning a specific test the teacher may feel reasonably safe in selecting it for use:

1. Does this particular test measure the abilities, habits, skills, attitudes, and information which I wish to measure?
2. How much time does it take to give the test? Is it long enough to give a reliable and consistent measure? (An unreliable measure is worse than useless.)
3. Is it easily and accurately scored?
4. Has it been widely used elsewhere?
5. Does it furnish accurate and extensive norms and standards for comparison and interpretation?
6. Is the interpretation of the scores simple and clear?
7. Do the results point the way to a remedial program?
8. Is the test economical in terms of time and money cost per unit of reliable information furnished by it?

Administration of the Test. One of the distinctive features of standardized tests is that they must be given under conditions which closely approximate the conditions under which they were standardized if the results are to be meaningful. Accordingly the teacher should follow the directions furnished with such tests. The attitude and personality of the examiner is also an important item in the administration of a test. The whole purpose of the test is defeated if an unnatural response is obtained. In the giving of any type of test care must be exercised to secure the cheerful confidence of the pupils.

Scoring the Test Papers. Much of the value arising from the use of tests in the classroom is lost if the teacher does not score the papers. In the first place the act of scoring the papers forces a closer inspection of their content than would otherwise be the case. Such careful scrutiny often reveals hitherto unsuspected features in the test. In the second place the teacher is brought into much closer contact with the pupil's special abilities and limitations by the act of personally scoring his test paper. For the majority of the better tests the task of scoring the papers is greatly simplified and objectified by the use of answer keys and scoring stencils.

The next step in the use of tests in the classroom is the summary and interpretation of the results.

306. Summarizing and Interpreting the Results of Testing. Skill in summarizing and interpreting test results is dependent upon the mastery of the following statistical techniques:

1. A knowledge of how to classify and tabulate data.
2. A knowledge of how to find the common measures of central tendency.
3. A knowledge of how to express the variability of data.
4. A knowledge of how to determine the relationship between two or more groups of data.
5. A knowledge of how to derive and use norms, standards, and derived scores for purposes of comparison and interpretation of test results.
6. A knowledge of how to present data in graphic form effectively.
7. A knowledge of the techniques for objectifying the marking system.

In addition to the use the classroom teacher may make of these skills in the proper utilization of test scores and in the assignment of grades there is the application which may be made of them in the study of current educational literature. Reports of progress in education are filled with statistical terms and techniques. The teacher can scarcely hope to keep abreast of the times in her profession if she is unable to read current educational literature understandingly.

307. **Diagnostic Testing and Remedial Teaching.** The analysis, identification, and measurement of the abilities which underlie and condition educational achievement unquestionably constitute the high point of the use of tests in constructive supervision. Forming the background of practically all possibilities of learning is that curiously interwoven maze of inherited traits, tendencies, and predispositions known as mental ability. Naturally enough, instruments designed to sample into this field constitute an important unit of the teacher's diagnostic equipment.

Mental Ability. The acceptance of the definition of mental ability as the power of the individual to learn, plus those informations, attitudes, and skills which he has developed through reacting to his environment makes it relatively easy to set up devices for its measurement and interpretation. Mental tests are incapable of securing a direct measure of capacity unaffected by experience and training. They do not measure the actual process of learning, or the quality of the learning equipment directly, but provide the basis for inferences as to the equipment from the amount of learning which has taken place under certain environmental conditions. The value of the mental test lies in the opportunity it affords for making this inference on a reasonably objective basis. Thus the mental test, care-

fully used and critically interpreted, constitutes a most effective and useful instrument for classroom diagnosis. Not only do mental test scores secure valuable evidence as to basic or general limitations and superiorities but they frequently offer most helpful hints as to the existence of more highly specialized abilities or disabilities. In the last analysis, prognostic tests which render such important service in certain types of educational and vocational guidance are specialized tests of intelligence.

Tool Subjects. Among the most fundamental of all school outcomes is the development of skill in the use of the tools of expression and communication as represented by such school subjects as composition, language, usage, spelling, handwriting, and reading. Perhaps arithmetic may even be included in this list. In theory, at least, these tool skills are mastered in the elementary school and do not constitute measurement problems for the junior and senior high school teacher. In practice this is not the case, for these subjects are rather notorious for the trouble they cause high school and even college students.

The chief function of the language skills is to make possible the inter-communication of ideas. This appears to be a simple enough process. A real mastery of the techniques by which the transmission of ideas takes place is so complicated, however, that most of us never attain it. Much remains to be accomplished by educators in the more critical analysis of tools of expression and communication, and in the construction of more adequate measurement and remedial instruments.

Accomplishment in almost all school subjects, but especially in the content subjects, is definitely conditioned by the pupil's ability to read. While reading does not permit analysis into numerous specific skills as may be done in arithmetic, considerable progress has been made in recent years in securing reasonably adequate analytical measures of reading disabilities, and in developing remedial procedures which are quite effective at the high school and college level.

Arithmetic is one of the tool subjects which not only has a very definite social utility but includes many skills themselves basic to successful work with more complicated mathe-

matical concepts. Due to the character of the organization of the subject matter of arithmetic, much progress has been made in the identification of its basic skills and the subsequent development of diagnostic and remedial materials.

Spelling and handwriting, two of the very important mechanical elements in written language, have both been analyzed and measured with reasonable success. The course of study content in spelling is becoming quite well known. The importance of providing the student with an adequate technique for learning to spell is generally recognized. The functions of testing in spelling instruction are apparent. The fact that by far the larger portion of testing in spelling is done in connection with teaching may ultimately carry over into other subjects.

Handwriting is a product of classroom and individual activity rather than the result of a pupil's response to stimulus. Products generally require evaluation by means of scales or score cards, and as a result the values assigned are frequently highly subjective. However, writing has responded to analysis sufficiently to enable the classroom teacher and supervisor to set up a rather definite corrective program in spite of the subjective elements involved in its measurement.

Content Subjects. Mental abilities and the special tool subjects lend themselves reasonably well to analysis and diagnosis. The content subjects, such as the social sciences and the more exact sciences, because of vagueness in the statements of their aims and purposes, are extremely difficult to measure in an analytical manner. Mastery of the tool subjects, such as reading, language, etc., enters into the measures of achievement obtainable in these fields to such a degree that many times the results are practically meaningless. However, as more critical analyses are made of curricular materials in the content subjects, and the aims and outcomes of instruction are more exactly stated, measurement in these fields is certain to be made more accurate and analytical.

Fine Arts. Changes in educational emphasis from the vocational and practical to the cultural, and increases in opportunities for the enjoyment of leisure are combining to force increased attention to the fine arts. That public

interest in good music and in the applied forms of art is being rapidly awakened is indicated by the many state and sectional music contests, and in the numerous community art projects now under way. The progressive classroom teacher of these subjects will wish to utilize all of the suggestions as to content, method, and outcomes which the tests in these fields make available.

General Educational Achievement. While the emphasis throughout this volume is definitely upon the measurement of the specific rather than the general aspects of school accomplishments, there is a recognizable need for the latter type of measurement. For general survey purposes, for evaluation of curricular content, and for later individual detailed diagnosis, such general achievement tests are valuable. The important point which the teacher should keep in mind, however, is the fact that such measurement is only the beginning not the end of classroom testing. Throughout all of the supervisory and instructional uses of educational tests it must be remembered that vague and indefinite measurement is useless; measurement which stops short of identifying the existence and causes of individual difficulties is futile; and that correction of defect without exact diagnosis is practically impossible.

308. Measurement and the Total Individual. There are certain aspects of ability, accomplishment, skill, aptitude, character, and personality which unquestionably lend themselves to reasonably objective measurement. The emphasis which is given to these measurable qualities frequently gives the impression that they represent the major elements in the total understanding of the individual. Such is far from the case, however, for many of the intangibles of the child's personality are almost certainly of greater importance, although in many cases they are practically impossible to measure objectively. This merely means that the teacher must be made keenly aware of the fact that something lies beyond objective measurement. He must see that appraisals of the child's total personality are basic to effective classroom teaching. He must recognize that many (probably most) of these vital appraisals must be made on the basis of keen observation and sympathetic analysis of his

pupils. Even if the teacher were gifted with unusual observational and analytic power, superior native capacity, natural sympathy; even if he were a four-year college or normal school graduate with graduate degrees in pedagogy, psychology, sociology, psychiatry and medicine, he could not hope to comprehend more than a few of the more superficial aspects of the child's personality. The important point here is that while it is impossible to know all, it is not impossible for the teacher to be cognizant of and sensitive to these problems.

309. **After Testing, What?** This question is in the back of the mind of every classroom teacher and every supervisor who has used current standardized tests. Much of the early use of tests was futile due to the fact that such broad vague phases of educational achievement were tested that even though reliable results were obtained, nothing specific could be done about the situation. Furthermore a great deal of the early use of tests in the classroom was a matter of satisfying curiosity. Hundreds of thousands of dollars' worth of tests have been given to millions of children, taking thousands of hours of pupil and teacher time, with absolutely nothing ever coming of it. Teachers have a right to expect that something tangible will be given them in return for interruptions and for pupil time spent in testing. One way to insure this return is for the teachers themselves to take an active part in the program. A type of training, an attitude toward their profession, a clearer insight into the difficulties faced by their pupils, is thereby gained which may not come to them in any other way.

The results of supervisory tests given periodically for the purpose of checking the efficiency of the teacher's instruction should be revealed to the classroom teacher in terms of specific suggestions for the further improvement of the situation. Instructional and diagnostic tests used by the teacher in the classroom should furnish such specific information concerning the abilities and limitations of the pupils that a program of preventive and corrective instruction can be begun at once.

In spite of the fact that a great deal of attention is now being focused on the problems of developing corrective and

remedial materials to use for the follow-up program after testing, only a limited amount of such material is now available. It is noticeable also that the main bulk of such material is limited to two or three subjects, namely, arithmetic, language, and reading.

310. **Conclusion.** Within the past score of years tests and measuring devices in nearly all subject-matter fields have been developed. What the future of this movement will be no one can predict. In many ways the rapid development has been unfortunate for it has resulted in confusion on the part of the classroom teacher, the one who should profit most from the program. There is a danger of over-production without sufficient refinement. This is illustrated in the tendency to flood the market with more and more test devices, with only an indifferent and mostly inadequate attempt to answer the question of what to do after testing. The encouraging aspect of the situation appears in the fact that high school teachers themselves are better trained professionally. They are becoming better informed in the techniques of using and interpreting all kinds of educational tests and are demanding that the corrective and remedial aspects of objective supervision shall materialize in the classroom. Unquestionably the greatest developments in measurements in the next decade will be in the refinement of existing measuring instruments to the point that accurate and reliable information concerning the individual difficulties of each pupil in each particular type of achievement may be obtained. The corollary of this development is, of course, the preparation of valid instructional and remedial materials for use when the weaknesses of pupils are revealed. When such material is available in the major subject-matter fields, the program of measurement in education will be justified and the instructional problems of the classroom teacher will be greatly simplified.

APPENDIX

In the various sections of the *Appendix* will be found the following material supplementing the discussions and explanations in the main body of the text:

A. Publishers and Distributors of Tests Mentioned in this Volume.
B. List of Educational Test Materials Mentioned in this Volume.
C. Glossary.

APPENDIX A

PUBLISHERS AND DISTRIBUTORS OF EDUCATIONAL TESTS MENTIONED IN THIS VOLUME

1. American Child Health Association
2. American Council Co-operative Test Service, New York City
3. A. N. Palmer Company, Chicago
4. Barnes, A. S., and Company, New York
5. Brooklyn Training School for Teachers, Brooklyn, New York
6. Bruce Publishing Company, Milwaukee, Wisconsin
7. Bureau of Co-operative Research, Indiana University, Bloomington
8. Bureau of Educational Measurements and Standards, Kansas State Teachers College, Emporia
9. Bureau of Educational Research and Service, University of Iowa, Iowa City
10. Bureau of Educational Research, University of Illinois, Urbana
11. Bureau of Educational Research, University of North Carolina, Chapel Hill
12. C. A. Gregory, Inc., Cincinnati, Ohio
13. Carl Fischer, Inc., Chicago
14. C. H. Stoelting Company, Chicago
15. Courtis Standard Tests, Detroit
16. Detroit Board of Education, Detroit
17. Educational and Personnel Publishing Company, Washington
18. Educational Test Bureau, Minneapolis and Philadelphia

19. Farnum Press, Minneapolis
20. Ginn and Company, Boston
21. Houghton Mifflin Company, Boston
22. Iowa Child Welfare Research Station, University of Iowa, Iowa City
23. J. B. Lippincott Company, Philadelphia
24. J. L. Hammet Company, Cambridge, Mass.
25. Lafayette Printing Company, Lafayette, Indiana
26. Manual Arts Press, Peoria, Illinois
27. Marietta Apparatus Company, Marietta, Ohio
28. Playground and Recreational Association of America, New York
29. Philadelphia Board of Public Education, Philadelphia
30. Public School Publishing Company, Bloomington, Illinois
31. Russell Sage Foundation, New York
32. Scott, Foresman and Company, Chicago
33. Southern California School Book Depository, Los Angeles
34. Stanford University Press, Stanford University, California
35. Teachers College Bureau of Publications, Columbia University, New York
36. Turner E. Smith and Company, Atlanta
37. University of Texas, Austin
38. World Book Company, Yonkers-on-Hudson, New York

APPENDIX B

LIST OF EDUCATIONAL TESTS MENTIONED IN THIS VOLUME

The classified list of educational tests presented here represents only the tests mentioned specifically in the various chapters of this book. It in no sense represents a complete catalogue of standard tests. The list is arranged within each subject on an alphabetical basis. The authors accept no responsibility for the merit of any of these tests.

The numbers in parentheses, as (1), refer to and identify the publishers as listed in the foregoing list of publishers of educational test materials. The numeral appearing after the title refers to the chapter in the textbook in which reference is made to the test.

Algebra

Breslich Algebra Survey Tests, (30), XVI
Columbia Research Bureau Algebra Tests, (38), XVI

Douglas Standard Diagnostic Tests for Elementary Algebra, (12), XVI
Hotz First Year Algebra Scales, (35), II
Iowa Algebra Aptitude Test, (9), XII, XVI
Iowa Placement Examinations; Mathematics Aptitude, (9), XII, XVI
Iowa Placement Examinations; Mathematics Training, (9), XVI
Iowa Every-Pupil Tests in Algebra, (9), XVI
Iowa Unit-Achievement Tests in First Year Algebra, (9), XVI
Lee Test of Algebraic Ability, (33), XVI
Orleans Algebra Prognosis Test, (38), XVI
Schorling-Clark-Lindell Instructional Tests in Algebra, (35), XVI

Arithmetic

Brueckner Diagnostic Tests in Arithmetic, (18), XVI
Compass Diagnostic Tests in Arithmetic, (32), II, XVI
Compass Survey Tests in Arithmetic, (32), II, XVI
Courtis Series B Arithmetic Tests, (15), XVI
Kinney Scales of Problems in Commercial Arithmetic, (30), XVI
Knight-McClure Arithmetic Neatness Scale, (32), II
New Stanford Arithmetic Tests, (38), XVI
Newland Chart for Diagnosis of Illegibilities in Written Arabic Numerals, (30), XVI
Unit Scales of Attainment in Arithmetic, (18), XVI
Woody-McCall Mixed Fundamentals Tests in Arithmetic, (35), II, XVI

Art

Kline-Carey Drawing Scales, (35), XIX
Lewerenz Test in Fundamental Abilities of Visual Art, (33), XIX
McAdory Art Tests, (35), XIX
Meier-Seashore Art Judgment Tests, (9), II, XII, XIX.

Biology

Blaisdell Instructional Tests in Biology, (38), XVII
Cooprider Information Exercises in Biology, (30), XVII
Michigan Botany Test, (30), XVII
Oakes and Powers Test of General Biology, (35), XVII
Ruch-Cossmann Biology Test, (38), II, XVII

Chemistry

Glenn-Welton Instructional Tests in Chemistry, (33), XVII
Harvard (Gerry) Test of High School Chemistry, (20), XVII

Iowa Placement Examinations ; Chemistry Aptitude, (9), XVII
Iowa Placement Examinations ; Chemistry Training, (9), XVII
Powers General Chemistry Test, (38), II, XVII
Rich Chemistry Tests, (30), XVII

Civics and Government

Almack Test in American Civics and Government, (12), XV
American Council Civics and Government Test, (38), XV
Bear Test on the United States Constitution, (30), XV
Brown-Woody Civics Test, (38), XV
Burton Civics Test, (38), XV
Hill-Wilson Test in Civic Action, (30), XV
Hill Civic Attitudes and Civic Information, (30), XV
Iowa Every-Pupil Tests in American Government, (9), XV
Magruder-Chambers-Clinton American Civics and Government Test, (30), XV
Patterson Test on Study Exercises on the Constitution of the United States, (30), XV

Economics

American Council Economics Test, (38), XV
Iowa Every-Pupil Tests in Economics, (9), XV

English Composition and Grammar

Briggs Form Test in Language, (35), II, XIII
Clapp-Young Self-Marking English Test, (21), XIII
Columbia Research Bureau English Test, (38), XIII
Cross English Test, (38), XIII
Franseen Diagnostic Tests in Language, (12), XIII
Hillegas Scale for Measuring Written Composition, (35), II, XIII
Inglis Test of English Vocabulary, (20), XIV
Iowa Elementary Language Tests, (18), II, XIII
Iowa Grammar Information Test, (9), XIII
Iowa Placement Examinations ; English Aptitude, (9), XIII
Iowa Placement Examinations ; English Training, (9), XII, XIII
Kirby Grammar Test, (9), II, XIII
Leonard Diagnostic Test in Punctuation and Capitalization, (38), XIII
Lewis English Composition Scales, (38), XIII
Markham English Vocabulary Test, (30), XIV
Nassau County Supplement to the Hillegas Scale, (35), XIII
Poley Précis Test, (30), XIV
Pressey Diagnostic Tests in English Composition, (30), II, XIII

Shutte English Diction Test, (30), XIII
Tressler English Minimal Essentials Test, (30), XIII
Wakefield Diagnostic English Test, (12), XIII
Willing Composition Scale, (30), II, XIII
Wilson Language Error Test, (38), II, XIII

French

American Council Alpha French Test, (38), XVIII
American Council Beta French Test, (38), XVIII
American Council French Grammar Test, (38), XVIII
Broom-Brown Silent Reading Test in French, (33), XVIII
Columbia Research Bureau French Test, (38), XVIII
Handschin Modern Language Tests, French, (38), XVIII
Harvard (Twiggs) French Vocabulary Test, (20), XVIII
Henmon French Test, (38), XVIII
Iowa Placement Examinations ; French Training, (9), XVIII
Sammartino-Krause Standard French Tests, (30), XVIII

General Achievement

American Council Co-operative Tests, (2), XVII, XXII
Iowa High School Content Examinations, (9), XXII
Iowa Placement Examinations, (9), XII, XIII, XVI, XVII, XVIII
Lippincott-Chapman Classroom Products Survey Test, (23), XXII
Metropolitan Achievement Test, (38), XXII
Modern School Achievement Tests, (35), XVII, XXII
New Stanford Achievement Examinations, (38), II, XVII
Otis Classification Test, (38), XXII
Pintner's Educational Achievement Tests, (35), XXII
Pressey Technical Vocabularies of Public School Subjects, (30), XVIII
Progressive School Achievement Tests, (33), XXII
Public School Achievement Tests, (30), XXII
Sones-Harry High School Achievement Tests, (33), II, XXII
Unit Scales of Attainment, (18), II

General Science

Dvorak General Science Scales, (30), XVII
Glenn-Gruenberg Instructional Tests in General Science, (38), XVII
Odell Scales for Rating Pupil's Answers to Thought Questions in General Science, (10), XVII
Pieper-Beauchamp Tests in Everyday Science, (32), XVII

Ruch-Popenoe General Science Test, (38), II, XVII
Van Wagenen General Science Reading Scales, (30), XVII

GERMAN

American Council Alpha German Test, (38), XVIII
American Council on Education German Reading Test, (30), XVIII
Columbia Research Bureau German Test, (38), XVIII

GEOGRAPHY

New Stanford Geography Test, (38), XV
Unit Scales of Attainment in Geography, (18), XV
Wiedefeld-Walther Geography Test, (38), XV

GEOMETRY AND TRIGONOMETRY

American Council Trigonometry Test, (38), XVI
American Council Solid Geometry Test, (38), XVI
Bishop-Irwin Instructional Tests in Plane Geometry, (38), XVI
Breslich Geometry Survey Tests, (30), XVI
Columbia Research Bureau Plane Geometry Test, (38), XVI
Iowa Plane Geometry Aptitude Test, (9), XII, XVI
Lane-Greene Unit-Achievement Tests in Plane Geometry, (20), XVI
McMindes Plane Geometry Tests, (30), XVI
Orleans Plane Geometry Achievement Test, (38), XVI
Orleans Geometry Prognosis Test, (38), XVI
Perry Geometry Tests, (25), XVI
Renfrow Diagnostic Tests in Plane Geometry, (12), XVI
Schorling-Sanford Achievement Test in Plane Geometry, (35), II, XVI
Seattle Solid Geometry Tests, (30), XVI
Webb Plane Geometry Test, (30), XVI

HANDWRITING

American Handwriting Scale, (3), XIII
Ayres Handwriting Scales, (31), II, XIII
Thorndike Handwriting Scale, (35), II, XIII

HEALTH AND PHYSICAL EDUCATION

Athletic Badge Tests, (28), XX
Baldwin-Wood Weight-Height-Age Tables, (22), XX
Brace's Scale of Motor Ability Tests, (4), XX

Detroit Decathlon for Boys, (16), XX
Franzen Health Education Tests, (1), XX
Gates-Strang Health Knowledge Test, (35), XX
McCloy General Motor Ability Test, (22), XX
McCloy Motor Capacity Test, (22), XX
Orleans-Sealy Health Test, (30), XX
Philadelphia Public School Age Aim Charts, (29), XX
Rogers Strength Test, (35), XX
Wood-Lerigo Health Scale, (30), XX

History, United States

Barr Diagnostic Tests in American History, (30), XV
Bowman United States History Test, (30), XV
Columbia Research Bureau American History Test, (38), XV
Denny-Nelson American History Test, (38), XV
Ely-King History Tests, (33), XV
Gregory Tests in American History, (12), XV
Gregory-Owens Tests in Mediaeval and Modern History, (12), XV
Harlan Test of Information in American History, (30), XV
Harvard (Kepner) Background Test in Social Sciences, (20), XV
Iowa Every-Pupil Tests in United States History, (9), XV
Iowa General Information Test in American History, (9), II, XV
Junior American History Test, (38), XV
Odell American History Scales, (10), XV
Pressey-Richard Tests on Understanding of American History, (30), XV
Van Wagenen American History Scales, (35), XV
Witham Hall of Fame Test, (24), XV

History, World

American Council European History Test, (38), XV
Iowa Every-Pupil Tests in World History, (9), XV
Vannest Diagnostic Test in Modern European History, (7), XV

Home Economics

Engle-Stenquist Home Economics Test, (38), XXI
Frear-Coxe Clothing Test, (30), XXI
Illinois Food Test, (30), XXI
King-Clark Foods Test, (38), XXI
Murdock Sewing Scale, (35), XXI
Stevenson-Trilling Patterns Test, (30), XXI
Streeter-Trilling Food Preparation Test, (30), XXI

INTELLIGENCE

Haggerty Intelligence Examinations, (38), XI
Illinois Examinations, (30), XI
Kuhlmann-Anderson Intelligence Examinations, (18), XI
National Intelligence Tests, (38), XI
Otis Group Intelligence Test, (38), XI
Otis Self-Administering Tests, (38), XI
Terman Group Test of Mental Ability, (38), XI

LATIN

Godsey Latin Composition Tests, (38), XVIII
Henmon Latin Tests, (38), XVIII
Hutchinson Latin Grammar Scales, (30), XVIII
Iowa Every-Pupil Latin Test, (9), XVIII
Lohr-Latshaw Diagnostic Latin Tests, (10), XVIII
New York Latin Achievement Test, (38), XVIII
Orleans-Solomon Latin Prognosis Test, (38), XVIII
Powers Diagnostic Latin Test, (30), XVIII
Pressey Test in Latin Syntax, (30), XVIII
Stevenson Latin Vocabulary Test, (30), XVIII
Stevenson-Coxe Latin Derivatives Test, (30), XVIII
Tyler-Pressey Test in Latin Verb Forms, (30), XVIII
Ullman-Kirby Latin Comprehension Test, (9), XVIII
White Latin Test, (38), XVIII

LITERATURE

Abbot-Trabue Scales for Judging Poetry, (35), XIV
Analytical Scales of Attainment in Literature, (18), XIV
Carroll Prose Appreciation Tests, (18), XIV
Logasa-Wright Tests for Appreciation of Literature, (30), XIV
Satterfield Objective Tests in English Literature, (36), XIV
Stanford Test of Comprehension in Literature, (34), XIV
Van Wagenen Reading Scales in Literature, (30), II, XIV

MANUAL ARTS

Badger Mechanical Drawing Test, (30), XXI
Castle Mechanical Drawing Test, (26), XXI
Detroit Mechanical Aptitude Examination, (30), II
Fischer Mechanical Drawing Tests, (6), XII, XXI
Hunter Shop Tests, (26), XXI
Minnesota Mechanical Ability Tests, (27), XII
Nash-Van Duzee Industrial Arts Tests, (6), XXI

Newkirk-Stoddard Home Mechanics Test, (9), XXI
O'Rourke Mechanical Aptitude Tests, (17), II, XII
Stenquist Assembling Tests, (14), XXI
Stenquist Mechanical Aptitude Test, (38), II, XI, XXI
Wells Knowledge Test in Machine Shop, (26), XXI
Wells-Laubach Achievement Test in Mechanical Drawing, (26), XXI
Wells-Laubach Industrial Arts Tests, (26), XXI
Wells Knowledge Test in Woodworking, (26), XXI

Miscellaneous Tests

Iowa Every-Pupil Academic Tests, (9), XV, XVI, XVII, XVIII
Iowa Placement Examinations, (9), II, XII, XVI, XVII, XVIII
Pressey Technical Vocabularies in Public School Subjects, (30), XV, XVIII

Modern Languages

(See also French, German, Spanish)

Iowa Placement Examinations; Foreign Language Aptitude, (9), XVIII
Luria-Orleans Modern Language Prognosis, (38), XVIII
Wilkins Prognosis Test in Modern Languages, (38), XVIII

Music

Beach Standardized Music Tests, (8), XIX
Drake Musical Memory Test, (30), XIX
Kwalwasser-Dykema Music Tests, (13), XIX
Kwalwasser-Ruch Test of Musical Accomplishment, (9), XIX
Kwalwasser Test of Music Information and Appreciation, (9), II, XIX
Providence Inventory Test in Music, (38), XIX
Seashore Musical Talent Tests, (14), XII
Torgerson-Fahnestock Music Test, (20), XIX

Personality

Bernreuter Personality Inventory, (34), XII
Loofbourow-Keys Personal Index, (18), XII
Strong Vocational Interest Blank, (34), XII

Physics

Black and Burlingame Elementary Physics Tests, (20), XVII
Columbia University Research Bureau Physics Tests, (38), II, XVII

Glenn-Obourne Instructional Tests in Physics, (38), XVII
Hughes Physics Scales, (30), XVII
Iowa Placement Examinations; Physics Aptitude, (9), XVII
Iowa Placement Examinations; Physics Training, (9), XVII
Kilzer-Kirby Inventory Tests for Mathematics Needed in High School Physics, (30), XVII
Kirkpatrick-Greene Handbook of Objective Test Exercises in Physics, (30), IV, XVII
Lapp's Iowa Achievement Examinations in College Physics, (9), XVII
Stewart-Ashbaugh Physics Tests, (30), XVII

Prognostic and Aptitude Tests

Detroit Mechanical Aptitude Examination, (30), II, XII
Iowa Algebra Aptitude Test, (9), II, XII, XVI
Iowa Plane Geometry Aptitude Tests, (9), II, XII, XVI
Iowa Placement Examinations; Chemistry Aptitude, (9), II, XVII
Iowa Placement Examinations; English Aptitude, (9), II
Iowa Placement Examinations; Foreign Language Aptitude, (9), II, XVIII
Iowa Placement Examinations; Physics Aptitude, (9), II, XVII
Lee Test of Algebraic Ability, (33), XVI
Luria-Orleans Modern Language Prognosis, (38), XVIII
MacQuarrie Test of Mechanical Ability, (33), XII
McCloy Motor Capacity Tests, (22), XX
Minnesota Mechanical Ability Tests, (27), XII
Orleans-Solomon Latin Prognosis Test, (38), XVIII
O'Rourke Mechanical Aptitude Tests, (17), II, XII
Orleans Algebra Prognosis Test, (38), II, XVI
Orleans Geometry Prognosis Test, (38), II, XVI
Seashore Music Talent Tests, (14), II
Stenquist Mechanical Aptitude Test, (38), II, XII
Wilkins Prognosis Test in Modern Languages, (38), XVIII

Reading

Detroit Reading Tests, (38), XIV
Gates Silent Reading Tests, (35), XIV
Haggerty Reading Examinations, (38), V, XIV
Iowa Silent Reading Test, Advanced Examination, (38), XIV
Iowa Silent Reading Test, Elementary Examination, (38), XIV
Nelson-Denny Reading Tests for Colleges and Senior High Schools, (21), XIV
Poley Précis Test, (30), XIV

Sangren-Woody Reading Test, (38), II, XIV
Stanford Achievement Reading Test, (38), XIV
Thorndike-McCall Silent Reading Scales, (35), XIV
Unit Scales of Attainment in Reading, (18), XIV
Van Wagenen Reading Scales in Literature and History, (30), XV
Whipple High School and College Reading Test, (30), XIV

Science

Iowa Every-Pupil Tests in Science, (9), XVII

Spanish

American Council Alpha Spanish Test, (38), XVIII
Columbia Research Bureau Spanish Test, (38), XVIII
Contreras-Broom-Kaulpers Spanish Reading and Vocabulary Tests, (30), XVIII
Handschin Modern Language Tests, Spanish, (38), XVIII
Iowa Placement Examinations; Spanish Training, (9), XVIII

Spelling

Ayres Spelling Scale, (31), XIII
Buckingham Extension of Ayres Scale, (30), XIII
Iowa Spelling Scales, (30), II, XIII
Morrison-McCall Spelling Scales, (35), XIII
Seven-S Spelling Scales, (35), XIII
Simmons-Bixler High School Spelling Scale, (36), II, XIII

APPENDIX C

GLOSSARY

For the convenience of the reader who may be unacquainted with many of the terms used in a somewhat special sense in this book this list of definitions is presented.

ability. The power to produce ; the result of training operating on innate capacity.

accomplishment. Used in the sense of production, achievement.

accuracy. Ratio of number of exercises attempted to number correctly done.

A.Q. — the Accomplishment Quotient. A ratio expressing relative educational development to mental development.

achievement. Used in the sense of accomplishment, production. Also used as descriptive of tests measuring abilities or products acquired primarily from school training.

acquisitive skills. Used here to represent those tool skills which are essential in the mastery of the content subjects such as silent reading ability.

adjusted individual. One who is properly placed mentally, physically, socially, etc., with respect to his environment.

administrability. One of the criteria for selection of tests. It implies the ease with which the material may be used, scored, and interpreted.

age equivalent. The age score assigned to given point scores on tests provided with age norms.

age norms. The typical or expected achievement of subjects of various ages usually expressed as medians or as averages of actual performance of subjects of different age groups.

alternate response test. An objective test of the type which permits the subject to choose between two possible answers, one of which is right and the other wrong. See also *true-false tests.*

ambiguity. The quality of a test item which causes the reader to be left in doubt as to its real meaning and intent.

analysis. Used in the sense of taking apart or reduction in the process of identifying specific skills.

analytical test. A test which takes several different cross-sections of abilities related to total accomplishment in the subject, thus furnishing a basis for the analysis of underlying skills.

aptitude test. Same as *prognostic test.*

arithmetic mean. The most commonly used of the measures of central tendency. Commonly called the average.

array. A collection of data usually organized around a particular function.

articulation. Used almost synonymously with pronunciation. Also a tendency for the muscles of the throat and larynx to function in silent reading.

assumed mean. The mid-point of the class-interval which is taken as the point about which the correction is made in computing the arithmetic mean by the short method from a frequency distribution.

astigmatism. A defect in the lens of the eye which causes rays of light from certain points or directions to fall out of focus, thus resulting in a blurred image.

auditory. Having to do with the sense of hearing.

average. A generic term covering the measures of central tendency, but commonly used to designate the arithmetic mean.

battery of tests. A group of several tests, covering a number of different subject-matter fields which are usually administered together. Illustrated by the *Stanford Achievement Tests.*

best answer test. A type of multiple-response test.

capacity. Power to learn or to profit from training.

central tendency. A term corresponding to average, commonly applied to the arithmetic mean, median, and the mode.

chance factor. Errors in measurement which arise as a result of the operation of non-systematic factors, as guessing, etc.

chronological age. The number of years the pupil has lived; his life age. May be expressed in either years or months.

class-interval. A limited series of consecutive points used as a pigeonholing or grouping device in tabulation. Used somewhat synonymously with group interval or step.

classification. The process of assigning an otherwise unassigned pupil to the particular unit or grade in a school for which his abilities and training best fit him.

coefficient of contingency. A measure of relationship used in the critical analysis of test items, based upon a comparison of the frequency of cases found in each category with the frequency which would be found if the traits were completely unrelated.

coefficient of correlation. An index of straight line relationship which ranges in value from plus one (+ 1.0) through zero to minus one (— 1.0). There are a number of indices of correlation, many of which are called *coefficients.* The use of the term *coefficient of correlation* has become somewhat restricted and refers generally to the Pearson Coefficient obtained by the so-called product-moment method.

comparable measures. Measures which are expressed in terms of the same unit or with respect to the same point of origin.

completion test. A common form of the new type of examination, consisting of statements in which one or more important words have been omitted. The test calls on the subject (person tested) to supply the missing words.

composite score. A single value used to express the results obtained from the use of a number of different measures. This may be the mean of scores yielded by several tests after these scores have been changed into equivalent form.

comprehension. A score denoting the degree of understanding of an exercise or material read.

conditions. Factors operating to introduce variants into testing or experimental situations.

content subjects. Fields in which mastery consists mainly in the acquisition of informations and attitudes, as the social sciences, the sciences, etc.

control group. The group which in an educational experiment is not subjected to the influence of the experimental factor used on the experimental group. The results from the two groups afford a basis for the evaluation of the experimental factor.

correction. A term used to imply a remedy or adjustment. Also used in a somewhat technical sense in the computation of the arithmetic mean, etc.

corrective. Having for its purpose the remedying of observed defects or difficulties.

correlation. The relation between two or more measures of the same individuals or items.

correlation table. A two-way or double-entry table which shows the relation between pairs of measures of the same individuals or items.

criterion. The standard by which the validity of a test may be determined.

C-scores. A unit of measurement or of item difficulty which represents one-tenth of a quartile deviation rather than one-tenth of a standard deviation as is used in the *T-score.*

decile. One of a division of ten equal parts.

derived scores. Values obtained through the changing of form of certain test scores, as ages, quotients, percentiles, etc., for the purpose of making data from different tests and sources comparable.

deviation. The amount by which a score or other measure differs from the central tendency of such a series of measures. Thus, if a child has a score of 14 and the median is 10, his deviation is plus 4.

diagnosis. Exact identification and location of strengths or weaknesses.

difficulty. A quality which when present in a test exercise results in a large per cent of incorrect responses.

discrimination. The quality in a test which causes it to distinguish adequately between varying levels of ability.

dispersion. See *deviation.*

educational age. Somewhat synonymous with achievement age; however, it usually applies to a pupil's standing in a number of school subjects while achievement age frequently refers to a single subject.

E.Q. — Educational Quotient. The ratio between the chronological age and the subject or educational age as found by a standard test. Found by dividing educational age, as found by suitable educational tests, by chronological age. Both ages are reduced to months and the resulting quotient is multiplied by 100 to keep it in whole number form. The quotient represents the relative educational development of the subject.

educational test. A term applied to a test which measures a pupil's mastery of subject matter taught in school.

empirical. Based or founded upon experience, experiential.

equivalent form. A duplicate or equal form of a standard test which yields a score approximately equal to the original form.

error of grouping. A variable error introduced by the practice of placing together in class-intervals unlike measures in tabulation.

error of sampling. A common difficulty arising in educational measurement resulting from the use of a limited number of cases taken as being typical of a larger group.

essay type examination. See *traditional examination.*

exercise. A unit of work in a test governed by a specific set of directions.

expectancy. Used here as anticipation of personal standards of accomplishment for the individual pupil.

experiment. The process of controlling conditions so that a particular situation may be repeated with only one variable operating.

f. An abbreviation used in certain tabulations to mean frequency.

fd. An abbreviation standing for frequency-deviation, one of the steps used in finding the arithmetic mean from a frequency table. It is found by multiplying the step-deviation from the assumed zero by the number of cases found at that class-interval.

first quartile (Q). The point on a scale of values such that 25% of the measures are less and 75% are greater.

fore exercise. A preliminary or practice exercise for the purpose of giving the pupil experience with the particular test situation.

form. A term used to distinguish between two or more arrangements of similar but not identical test exercises. Usually two forms of a test are approximately equal in difficulty.

frequency. The number of measures in a given class interval of a frequency tabulation.

frequency curve. A graphic representation of a distribution of measures.

frequency table. A distribution showing the number of cases or measures at each class-interval when put in tabular form.

fulcrum. The axis upon which a lever is supported and rotated.

functional analysis. Analysis according to use.

general intelligence test. A test designed to measure general native capacity. Contrasted with achievement test.

general survey test. A test designed to measure a number of different aspects of a single school subject or a test composed of a number of parts, each covering a phase of a different school subject. Contrasted with diagnostic tests.

gradation. The process of assigning the individual pupil to the organized school group with which he may be expected to do his most efficient work.

grade equivalent. The grade or portion of the grade nearest which a pupil's test score places him when compared with the norms of the test.

grouping. The process of classifying or tabulating data into certain steps or categories.

group test. A test designed to be given to a number of different pupils at the same time.

G-scores. Practically synonymous with grade scores, or grade equivalents.

half-half correlation. The coefficient which results from the correlation of the odd-even scores on a test.

half-sum. A term used in connection with the finding of the median. It is equal to $N/2$.

handedness. The predisposition to use one hand rather than the other in manual reactions.

Heinis growth units. A mathematically derived series of units to be used in computing the Personal Constant (*P.C.*).

higher decade. The addition facts involving an addend of more than units place, as $22 + 7$.

hyperopia. A visual defect resulting in unduly acute distance vision with marked weakness in near vision.

imbalance. A defect in the muscles of the eye causing the eyes to fail to move together.

improvement. Used in measurements to designate growth in accomplishment resulting from a definite interval of training.

individual differences. The observed or measured differences between individual pupils in capacity, ability, industry, etc.

informal test. A test prepared by the classroom teacher as distinguished from the standardized test.

instructional test. Tests designed to be used directly in connection with the teaching of a unit of material. Content commonly parallels closely the instructional unit.

I.Q. — the Intelligence Quotient. An expression of relative brightness. Found by dividing mental age as obtained from an intelligence test by chronological age. Both ages are usually reduced to months and the quotient is pointed off as a whole number.

intelligence. Defined most simply and practically as the power to learn or to profit from training.

interval. A limited series of scale points set up for grouping or pigeonholing data in tabulation. Used somewhat interchangeably with "step."

interpolation. The process of locating an intermediate point between two known points in accordance with the laws operating in the case at hand.

interpretation. Used mainly in the sense of explaining results of test scores and applying them to concrete or specific situations.

inventory test. A test designed for use as a preliminary check on the degree of mastery existing prior to instruction.

item. Used almost synonymously with test exercise.

manuscript writing. A free-hand lettering process in which the letters are not connected as in common script writing. Used widely in the primary grades.

marking system. The program by which the teacher's evaluation of the pupil's accomplishment is recorded.

mean. Usually signifies the arithmetic mean ; the point on the scale about which the deviations are zero.

median. A common measure of central tendency. A point on the scale such that there are as many equal to or larger than, as there are equal to or smaller than. The middle of a series of scores with these scores arranged in order of magnitude.

mental ability. Used somewhat synonymously with intelligence ; power to learn.

mental age (M.A.). The mental ability of a subject expressed in terms of the age of an average child having that ability. Thus, if a child has a mental age of an average 10-year-old child he is said to have a mental age of 10 years.

merit scale. A series of products arranged in definite order of quality for use in evaluating the quality of similar products.

mid-measure. The middle measure of a series of values arranged in order of magnitude. Similar to the median.

mid-point. The exact middle of a step or class-interval in a frequency table. Used as the point about which to correct in finding the arithmetic mean by the shorter methods.

multiple-choice test. A test composed of exercises arranged in such a way as to require the pupil to select one or more correct responses out of a group of several possible responses.

myopia. A visual defect resulting in unduly acute near vision with marked weakness in far vision.

N. A symbol commonly used to designate the number of cases in an array.

negative correlation. A relationship in which the larger values of one variable are accompanied by the smaller values of the other variable.

normal. Typical ; making regular progress, growth, or development.

norms. The median or average performance of subjects of different ages or grades, as determined by the testing of large numbers of cases.

objective. A term used in describing tests in which no opportunity for disagreement as to correctness of response exists.

odd-even scores. The scores which result from determining the number of correct responses to the odd-numbered and even-numbered exercises in a test.

opposites test. A test consisting of a list of terms calling for the opposite meaning of each to be given.

percentile. The points which divide the total number of cases in a given frequency distribution into 100 equal parts.

performance. Accomplishment, achievement. Used also to distinguish test scores as such from capacity.

Personal Constant (P.C.). A proposed substitute for the *I.Q.* which appears to reflect relative mental development without the operation of certain mathematical distortions due to the character of the mental growth curve.

phonation. The process of uttering articulate sounds.

power tests. Devices which measure achievement in terms of the difficulty of the task which the subject is just able to perform.

practice effect. The effect which previous experience with a test has on a later trial.

précis. A summary.

preventive teaching. Steps taken in connection with initial instruction to protect against the later appearance of specific weaknesses.

prognostic test. A test having for its purpose the prediction of probable future achievement based upon present performance.

quality scales. Devices, such as the *Ayres Handwriting Scales*, which measure by comparison with a set of standard specimens the outcome of the application of some specific skill.

quartile deviation. One-half of the scale range of the middle half of the cases in a distribution.

quartiles. The results of dividing a distribution into quarters. The first quartile is the lower one-fourth, the second quartile is the median, the third quartile is a point equalled or excelled by only one-fourth of the cases or measures.

quotients. Different types of ratios designed to reveal in a single numerical index the relative development of the individual with respect to the norm.

random sample. The process of selecting a limited number of cases from an unlimited group without allowing personal bias to enter into the selection. A random sample is one selected on a purely chance basis.

range. The limits of the scale necessary to include all measures in a distribution. Found by taking the smallest measure in the array from the largest.

rank. Position assigned to a score in a series.

rate score. A score expressing a pupil's rate of work.

rate tests. Devices which measure achievement in terms of the number of tasks of uniform difficulty which can be performed in a specified time.

reading readiness. The predisposition on the part of the child to profit from training in reading. The result of natural aptitude plus certain types of preliminary exercises and experiences.

recall test. A new type of test exercise so stated as to be answered correctly by the use of a single word contrasted with recognition tests.

recognition test. A test containing exercises of the true-false or multiple-answer type. These exercises do not force the student to formulate the answer.

relative rank. Position assigned cases or scores in a limited array.

reliability. Applied in a somewhat technical sense to tests as a means of expressing the consistency of performance of the test itself.

remedial. Material and devices having for their purpose the correction and elimination of observed difficulties and weaknesses in pupil accomplishment.

retarded. A term applied to an individual whose school progress, mental or educational development, is slower than is expected of the normal subject.

right-minus-wrong formula. Formula for correction of chance in scoring alternate-response exercises.

sampling. See *random sample.*

scale. This word is used in two different senses in measurement. In the narrow sense "scale" means the portion of every measuring instrument which is used to express the pupil's performance. Scale is also used in a general sense to designate a measuring instrument in which the scale for expressing performance is the distinguishing characteristic.

scores. A description of the performance of a subject.

sectioning. Placement of pupils in more or less homogeneous groups for instructional purposes. Usually involves a much finer classification than is involved in gradation.

sigma. The common abbreviation for standard deviation.

sigma units. The measuring unit which results from turning the measures comprising the scale into the relations they bear to the standard deviation of the distribution.

smoothed data. Data representing too few measures to be truly representative of a situation often show irregularity not typical of the entire group. The so-called process of smoothing, irons out these irregularities and approximates the type of distribution which would result if the sample included an infinite number of cases.

social sciences. A general group of content subjects including civics, government, economics, history, etc.

social utility. A modern point of view in the selection of the curricular material which holds that all subject matter, to warrant a place in the course of study, must contribute definitely to the training of the child to fit into adult society.

standards. The level of achievement in a given subject to which a school should bring its pupils. Norms are usually medians or averages based on actual achievement of subjects while standards are goals to be attained.

standard deviation. One of the more common measures of deviation or variability of scores.

standardized tests. A test is called *standardized* when the exercises comprising it have been properly selected and evaluated, and when in addition norms or standards have been established for it.

standardization. The process of setting up suitable objective goals of performance for tests or other instructional materials.

step. A term used interchangeably with class-interval. Refers to interval in a frequency tabulation.

stereopsis. The quality of depth in vision, three-dimensional perception.

sub-total. A term used in connection with the finding of the median.

subjectivity. The degree to which measurement results are affected by personal factors or judgments. Objectivity and subjectivity are both relative terms. All tests are somewhat subjective ; some are more so than others.

survey tests. Used to describe tests which have for the main purpose the measurement of abilities in terms of broad general functions.

synthesis. The process of putting together in an effective unit underlying and somewhat isolated skills so that they form an effective unit.

tabulation. The process of grouping or classifying data for purposes of condensation and interpretation.

talent. Used here as synonymous with aptitude.

teacher's marks. Numerical or other expressions of the teacher's evaluation of the pupil's accomplishment in a special field of activity.

technique. As applied to testing it commonly means a specific procedure.

test. Used in a general sense to describe any type of instrument for measuring any mental ability. More specifically it is the portion of a measuring device by which the pupil's performance is secured.

traditional examination. Refers to the type of examination commonly used previous to the development of the so-called objective-type tests. In the main this type consists of discussion exercises.

training. Any influence of environment or special educational program.

transfer of training. The utilization of common elements in two learning situations to simplify the learning process.

true-false test. A form of recognition test in which the student is called upon to determine the truth or falsity of the exercises.

T-scores. Derived scores based upon the standard deviation unit of the distribution.

ungrouped. Used here in contrast with grouped or classified in class-intervals.

unit scales. Tests in different subject-matter fields which are calibrated and equated in terms of a comparable unit of measurement.

validity. A term used to express the degree to which a measuring instrument measures the thing it purports to measure, or the extent to which it measures a thing which the subject tested has had an opportunity to learn.

variable. Refers to a quality which may exist in different amounts.

variability. Synonymous with *deviation.*

work-type. Term used in describing the silent reading skills commonly utilized in study.

yes-no tests. A variety of alternate-response tests comparable to the true-false type.

INDEX

A

Ability, defined, 210-211
Absolute ranks, 203-204
Accomplishment quotient (A.Q.), 241-243
Achievement age, 223-228
Achievement tests, kinds, 24-26; uses of, 91-96
Achievement tests, industrial education, 526-540
Acquisitive Skills, 371
Adams, R. M., 443
Adjustment, through measurements, 292-307, 557-562; and guidance, 282; school responsibility for, 283-284; as related to personality, character, 284
Administrability, 113-115
Administering tests, 122
Age-norms, interpretation of, 228-230
Age scores, as derived scores, 235-238
Alexander, H. B., 469
Algebra, aptitude, 402-405; measurement in, 397-405; objectives, 397; tests in, 27, 398-405
Alienation, coefficient of, 198
Allen, C. R., 542
Allen, R. D., 308, 555
Almack, John C., 442
Ambiguity in items, 136
American Child Health Education Tests, 500
American Classical League, 447
American History Tests, 376-379
Analytical diagnosis, 36-38
Anderson, L. D., 308
Anderson, Rose, 266, 280
Aptitude, measurement of, 21-24, 402-405
Arithmetic, diagnosis of, 394-397; skills, 391-394; tests, 394-397
Arithmetic mean, classroom use of, 164-165; definition of, 161; illustration of calculation of, 164; summary of steps in calculation of, 164
Art, instruction, outcomes of, 481-483; measurement of art judgment, 305, 484-486; measurement of accomplishment in, 486
Artistic Talent, 484-486
Art Judgment, measurement of, 305, 484-486
Ashbaugh, E. J., 347
Athletic Badge Test, 515-517
Averill, L. A., 350
Ayres, L. P., 8, 342-347
Ayres Handwriting Scales, 223, 342

B

Badger, A. J., 540
Badger Test in Mechanical Drawing, 69
Baldwin, Bird T., 519
Baldwin-Wood Weight-Height-Age Tables, 504-505
Ballenger, H. L., 141, 258, 324, 347
Bangs, C. W., 184, 571
Banker, H. J., 57, 88
Barton, William A., 387
Basis for testing program, 103-105
Bawden, W. I., 308
Beard, Chas. A., 387
Beauchamp, W. L., 442
Bennett, C. A., 540
Benson, C. E., 285
Bernreuter personality inventory, 295-297
Bernreuter, R. G., 295, 308
Betts, E. A., 315, 348, 367
Bibliographies, 8, 39, 57, 88, 98, 126, 148, 208, 245, 257, 280, 308, 347, 367, 387, 415, 442, 469, 490, 519, 540, 555, 571
Binet-Simon, 263-264
Biology, aims in, 435; tests in, 436-438
Bixler, H. H., 18, 555
Blackhurst, J. H., 469
Blanton, Smiley, 316
Bolton, F. E., 57
Book, W. F., 350, 367
Bovard, John F., 520
Brace, D. K., 508, 520
Breed, F. S., 347
Breslich, E. R., 415
Breslich Algebra Survey test, 399
Brewer, John M., 308
Briggs, T. H., 540
Brinkley, Sterling G., 61
Brooks, F. D., 490
Brownell, C. L., 494, 521

Brownell, H., 442
Brown-Woody Civics Tests, 376
Bruce, H. W., 294, 308
Brueckner, L. J., 39, 98, 126, 347, 367, 387, 415, 469, 556, 571
Buchanan, M. A., 469
Buckingham, B. R., 88, 245
Bureaus of research, 4-5
Burgess, T. O., 442
Burton, Richard, 347
Burton, W. H., 495
Buswell, G. T., 368, 415

C

Caldwell, O. W., 42, 57
Capacity, defined, 210-211
Carey, Gertrude L., 490
Carter, H. D., 309
Cassidy, R. R., 521
Castle, D. W., 540
Cattell, Psyche, 280
Central Tendency, measures of, 160-170
Cerf, Barry, 469
Chance factor in objective tests, 84-87
Chapman, J. C., 538, 541
Character traits, measuring, 295-300; rating scales for, 295-300
Chemistry, measurable qualities in, 433; objectives in, 432
Civics and Government tests, 381-384
Clark, W. W., 556
Class grades, use of sigma for analysis of, 183-185
Class diagnosis, 270
Classical League, 469
Classification of pupils, 560
Class-intervals, 153, 155-156, 158
Cleeton, Glen U., 541
Cline, E. C., 347
Coefficient of Correlation, meaning of, 189-190, 196-200; uses of, 198-200
Cole, Robert D., 446, 469
Cole-von Borgersrode scale for rating tests, 117-120
Coleman, Algernon, 446, 469
Columbia Research Bureau Algebra Test, 252
Columbia Research Bureau Physics Test, 31, 428
Compass Arithmetic Tests, diagnostic, 33-34, 229-230, 395; survey, 394
Commercial Arithmetic tests, 396
Completion test items, 80
Composition, measurement of, 319-322
Conner, W. L., 555
Contingency, coefficient of, 138-140
Cook, W. W., 332
Correction for guessing, 84-87
Correlation, computation of Pearson Product-Moment, 191-196; coefficient, meaning of, 196-198; meaning of, 189-190; table, 191; uses of, 198-200
Courtis, S. A., 42, 57, 98, 259, 321
Cozens, F. W., 520
Craig, G. S., 442
Cram, Fred D., 221
Crawford, J. R., 217
Criteria for tests, 105-116
C-scores, 237-238
Cunningham, H. A., 442
Curtis, F. D., 88, 443

D

Dansdill, Theresa, 520
Davis, G., 347
Dearborn, W. F., 280
Deciles, 204
DeGraff, M. H., 86, 89
Department of Superintendence, 443, 490
Derived scores, 231-243
Deviation, *See variability*
Diagnosis, basis for remedial work, 251-253; basis for preventive work, 254; function of, 248-256; meaning of, 247; nature of, 249
Diagnostic tests, vs. analytic tests, 32-38; in Arithmetic, 395
Dictation Exercise, 336
Difficulty of exercises, 73
Directions for objective tests, 70
Disciplinary uses of tests, 559
Discriminative Power, 138
Discussion exercises, 7, 38, 56, 87, 98, 124, 147, 257, 279, 307, 346, 367, 386, 414, 442, 468, 488, 519, 540, 555
Donson, George C., 541
Double-entry or correlation table, 191
Double true-false exercises, 84
Douglas Diagnostic Algebra Tests, 252, 253
Downing, E. R., 443
Drake, R. M., 477
Drake Musical Memory Tests, 477
Dykema, P., 490
Durrell, D. D., 368

E

Economy in testing, 115
Eddy, Helen M., 445, 469
Edgerton, A. H., 309, 541
Educational age, 223-230, 235-238
Educational diagnosis, 247-251
Educational quotient (E.Q.), 240
Educational tests, characteristics of, 10; classification of, 13-38; criteria for, 105-116

English, importance of, 311; objectives of, 312; measurement of oral, 315-319; measurement of written, 319-328; tests, 319-328
Error of grouping, 154
Essay-type tests, 10; functions of, 43-46; improving, 47; scoring, 53; types of exercises, 52
European History tests, 380-381
Evaluation of tests, 105-116
Examinations, characteristics of informal, 58-62; functions of, 43-46; improvement of, 47; new type, 59; significance of, 46; traditional, 47-56
Examination questions, score card for, 52
Exercises, arrangement of, 73-75; classification of, 79; criteria for, 134-140; difficulty of, 75; form of statement, 134; number of, 70; objective types, 79-84

F

Fauber, O. W., 54
Fischer, Ferdinand A. P., 541
Fischer Mechanical Drawing Tests, 528
Fisher, Rev. George, 2
Flaherty, E. B., 541
Flam, August, 541
Foreign Language, measurement in, 447-468; shifting objectives of, 445-447; tests in, 448-454
Forms of tests, equality of, 140-142
Foster, R. R., 88
Franzen, R., 241, 245, 275, 520
Freeland, G. E., 443
Freeman, Frank S., 387, 556
Freeman, F. N., 280, 343, 344, 347
Freeman Diagnostic Handwriting Scale, 339
French, achievement in, 457; tests, 457-462
Frequency table, steps in making, 152-153

G

Garrett, H. E., 181, 208
Gates, A. I., 337, 368, 556
Gates-Strang Health Test, 499-500
General Achievement Batteries, advantages and disadvantages, 544-547; types of, 547
General science, objectives in, 423; tests, 29-30, 423-426
Geography tests, 383
Geometry, measurement in, 405-411; objectives, 405; prognosis, 411-412
Germane, C. E., and E. G., 309
Gilliland, A. R., 387, 556
Glenn, M., 491
Good, Warren R., 208
Goodenough, Florence L., 490
Gradation of pupils, 95
Grade norms, interpretation of, 228-230; 232-235
Grading system, 182-185, 566-571
Graham, F. B., 555
Grammar tests, 322
Graphic record card, 236
Gray, W. S., 258, 349, 352, 368
Greene, H. A., 77, 88, 126, 141, 184, 208, 221, 223, 246, 258, 292, 315, 324, 347, 348, 415, 541, 571
Greene, Charles E., 415
Gregory, C. A., 251
Gregory American History Tests, 378
Gregory-Owens Medieval and Modern History test, 380
Grouping of test scores, 151-158
Group tests of mental ability, 265
G-scores, 232-234
Guessing in objective tests, 84-87
Guidance, 96, 286-295, 557-559

H

Hahn, H. H., 88
Hall, F. H., 443
Hamilton, E. R., 57
Hammond, H. P., 443
Handwriting, diagnosis of, 343; measurable qualities, 338-340; remedial work, 343-346; scales, 342-343
Hartshorne, H. A., 309
Health Education, corrective work, 517-519; objectives of, 494-496; preventive work, 517-519; significance of, 493; tests, 499-500
Heinis, H., 277, 278, 280
Heinis mental growth units, 277
Henmon, V. A. C., 456, 469, 470
Herring, J. P., 280
Hilden, A. H., 277
Hildreth, Gertrude H., 555
Hillegas, Milo B., 319
Hillegas Composition Scale, 319, 321
Hilpert, R. S., 491
History tests, types of, 373-386
Hjerstedt, W. G., 541
Hoke, K. J., 39, 89, 99, 126, 246, 348, 388, 520, 556, 571
Hollingworth, Leta S., 337
Holzinger, Karl J., 88, 208
Home Mechanics Test, 532-534
Horn, Ernest, 18, 348, 352, 369, 387
Hotz Algebra Scales, 27
Household Arts, tests in, 534-536
Hull, C. L., 309, 541
Hunter, W. L., 541
Hunter Shop Tests, 531-532

Hurd, A. W., 443
Hurt, A. O., 342

I

Index, ability to use, 357
Individual mental examinations, 263-265
Individual pupil diagnosis, 271
Industrial Arts tests, standardized, 530
Industrial Education, aptitude and intelligence tests in, 536-537; defined, 522; measurable factors, 526-527; objectives of, 524-525; tests in, 526-540
Informal Objective Tests, advantages of, 62-66; difficulty of items in, 73, 75; disadvantages of, 67-69; measurement by, 564-571; rearranging items in, 73, 75; samples of, 79-84; securing objectivity, 62; securing reliability, 64-66; securing validity, 72; steps in building, 70-77
Informal vs. standardized tests, 60-62
Instructional tests, algebra, 402; biology, 437; chemistry, 435; general science, 426; geometry, 410; physics, 429-430
Intelligence, defined, 260; measurement of, 260-262; methods of measuring, 261; tests of, 263-270
Intelligence Quotient (I.Q.), 239, 275-276
Inter-dependence of mental and achievement tests, 273-275
Interpretation of tests, 101 ; 577-578
Iowa Academic Tests, 384-385; 440; 553-554
Iowa Algebra Aptitude Test, 22-24, 306, 403-404, 561-562
Iowa Elementary Language Test, 141, 252, 324
Iowa General Information Tests, 30-31, 375
Iowa Grammar Information Test, 322
Iowa High School Content Examinations, 552-553
Iowa Plane Geometry Aptitude, 307, 411-412
Iowa Silent Reading Tests, 144, 217, 225, 231, 356-358
Iowa Placement Examinations, 403, 431, 434, 467
Iowa Revision of Brace Test, 508-510
Item, defined, 70
Items, difficulty of, 73; rearranging, 73-75

J

Johnson, Henry, 387
Jones, Arthur J., 286, 291, 309
Jordan, J. N., 470
Jordan, R. H., 387, 556
Judd, C. H., 390, 415, 490

K

Keene, Charles H., 520
Kefauver, G. N., 309
Kelley, T. L., 39, 57, 98, 126, 208, 246, 556
Kelley, V. H., 258
Kelly, F. J., 54
Kelsey, F. W., 470
Kerr, James, 520
Keys, N., 297, 309
Keyser, C. J., 390
Kilzer, L. R., 443
Kilzer-Kirby Inventory Test, 427
Kinney Commercial Arithmetic Scales, 396
Kirby, C. Valentine, 482, 490
Kirby, T. J., 452
Kirby Grammar Tests, 322
Kirkpatrick, J. E., 68, 77, 443
Kitson, Harry D., 541
Klar, W. H., 481, 490
Kline, L. W., 490
Kline-Carey Drawing Scales, 486
Knight, F. B., 126, 246, 258, 393, 415
Koos, L. V., 309, 340
Kramer, Edna E., 208
Krause, C. A., 470
Kuhlmann, F., 266, 267, 280
Kuhlmann-Anderson Intelligence Tests, 266-268
Kwalwasser, J., 472, 478, 491
Kwalwasser-Dykema Music Tests, 477
Kwalwasser-Ruch Musical Accomplishment Test, 478

L

La Brant, Lou A., 368
Lane, Ruth E., 415
Lane-Greene Unit-Achievement Tests, 407-408
Lang, A. R., 39, 41, 57, 88, 98, 126, 442
Langlie, T. A., 415
Latin, achievement tests, 448-453; possibilities of measurement in, 447-448; prognosis in, 453; remedial work, 454
Lessenger, W. E., 242
Lewerenz, Alfred S., 491
Lewerenz Art Abilities Test, 487, 489
Lewis English Composition Scales, 320
Lide, E. S., 415
Ligda, Paul, 416
Lindquist, E. F., 208, 387
Literature, objective tests in, 364-366; reading scales in, 366
Loofbourow, G. C., 297, 309

Loofbourow-Keys Personal Index, 297
Lough, J. E., 285
Lowell, A. L., 57
Lyman, R. L., 348

M

MacPhee, E. D., 469
MacQuarrie, T. W., 309
Madsen, I. N., 39, 98, 126, 292
Mann, Horace, 42
Manuel, H. T., 484, 491
Marking system, functions of, 566-571; objectifying the, 566-571
Matching exercises, 79, 81
Mathematics, aptitude, 306-307; in Junior-Senior H. S., 390-391; significance of, 389; tests in, 27, 398-411
Mathews, C. O., 387
May, M. A., 309
Mays, A. B., 541
McAdory, Margaret, 491
McAdory Art Test, 486
McAndrew, William, 57
McBroom, Maude, 352
McCall, W. A., 126, 148, 274
McCallister, J. M., 387
McCloy, C. H., 511, 512, 520
McClusky, H. J., 88
McCormick, C., 416
McKee, Paul, 348, 368
Mead, A. R., 126
Mean, arithmetic, calculation of, 160-164; definition of, 161; uses of, 164-165
Measurable skills, in algebra, 398; in arithmetic, 392-394; in fine arts, 472, 484-486; in foreign languages, 447-448; in geometry, 406; in handwriting, 338-340; in English, 315-318, 319-328; in reading, 352-355; in science, 421; in social studies, 371-373; in spelling, 328-329
Measurement, growth of interest in, 5; pioneers in, 2; significance of, 9-13
Mechanical Aptitude, measuring, 303-305
Mechanical Drawing tests, 527-530
Mechanical features of tests, 113-115
Median, calculation of, 170; definition of, 168; uses of, 167
Meier, Norman C., 306, 309, 484, 491
Meier-Seashore Art Judgment Tests, 484-485
Melby, E. O., 98, 126, 347, 367, 387, 415, 571
Mental ability, group tests of, 265; individual test of, 263; meaning of, 260; measuring, 263-266; nature of, 259
Mental age, 275
Mental tests, 20; kinds, 263; uses of, 96-97, 270
Mental test score, meaning of, 210-211, 260
Metropolitan Achievement Tests, 240
Mid-measure, calculation of, 167; definition of, 168
Mitchell, B. C., 208
Modern languages, measurable elements in, 456; prognosis in, 464; remedial work, 465-467
Modern School Achievement Tests, 240, 551
Monroe, M., 368
Monroe, W. S., 148
Morgan, B., 470
Morrison-McCall Spelling Scale, 333
Mort, P. R., 556
Morton, R. L., 416
Moss, F. A., 294, 309
Motor Ability, scales for, 510-513
Motivation of learning, 44
Multiple-response exercises, 79, 81-82
Mueller, A. D., 350
Mursell, J. L., 491
Music, achievement tests in, 475-481; aims and outcomes, 472-474; measurable qualities, 472-475; prognosis in, 475-477; remedial work in, 481; tests, 475
Musical Talent, 306; 475-477

N

Nash, Harry B., 541
Nash, J. B., 520
Nassau County Supplement to Hillegas Scale, 319
Newkirk, L. V., 134, 541
Newkirk-Stoddard Home Mechanics Test, 133, 532-534
New Stanford Achievement Tests, 228, 240, 547-548
New type tests, 59
Normal Curve, 178
Norms, computation of, 220-221; kinds, 215-221, 223-228; meaning of, 213, 221; reliability of, 145-147; standards and, 214; uses of, 214-215
Nyquist, F. V., 491
Nystrom, Ellen, 348

O

Objective examinations, advantages of, 62-66; characteristics of, 62-66; construction of, 70-77; disadvantages of, 67-69; using results from, 77-78, 564-571
Objective exercises, classification of, 78-79; samples of, 79-84

Objective tests, standardized and informal, 14, 20
Objectivity, 111-113
O'Brien, J. A., 351
Odell, C. W., 39, 57, 88, 98, 126, 148, 208, 246, 309, 348, 416, 443, 520, 541, 556, 571
Opdycke, J. B., 57
Oral English, diagnosis of, 315-319
Orleans, J. S., 88
Osburn, W. J., 53, 57
O'Shea, M. V., 470
Otis, A. S., 181; Self-Administering Test, 269
Otis Test-rating scale, 121

P

Paterson, D. G., 88, 89, 309
Patterson, A. S., 470
Percentile scores, 204-206
Performance, defined, 210
Personal Constant, 276-279
Personality and Character Traits, scales for, 295-300
Peters, C. J., 443
Phillips, C. A., 370, 387
Physical Achievement Tests, 513-517
Physical Capacity Tests, 506-509
Physical Education, measurable qualities, 497-499; objectives of, 497-499
Physical Growth, measurement of, 501-502
Physics, measurable qualities in, 427; tests, 427-432
Pierce, Anne E., 491
Pintner, R., 280
Piper, A. H., 294, 309
Powers, S. R., 432, 443
Power tests, 15, 25
Pressey, L. C., 372, 387, 468
Pressey Technical Vocabularies, 363
Preventive instruction, 254
Problem solving, tests, 374
Problems, 159, 160, 166, 167, 170, 171, 176, 187, 200, 201, 203, 207, 244, 245
Product Scales, 26
Prognostic tests, 21, 402-405, 453-454
Progressive School Achievement Tests, 551-552
Promotion and classification, 560
Providence Inventory Test in Music, 480

Q

Q, as Measure of Variability, 173-176
Quality Scales, 26
Quartiles, 175
Quotients, 239-243

R

Range, as measure of variability, 173; of scores, 153
Ranks, absolute, 203-206; assignment of, 201-206; relative, 201-203
Rate tests, 15, 24
Reading, abilities, 353-356; defects, 360-362; objectives, 351-355; tests, 356-360
Recall exercises, 80-81
Recognition type exercises, 81-84
Reeve, W. D., 416
Relationship, measures of, 187-201
Relative ranks, 201-203
Reliability, 108-110, 142-145; sampling as factor in, 109; of norms, 145-147
Research Bureaus, 4-5
Rice, J. M., 3
Rich, S. G., 444
Richards, O. W., 444
Riley, Gordon, 280
Rogers, F. R., 502, 506, 507, 520
Rowell, H. G., 521
Ruch, G. M., 39, 48, 57, 58, 72, 86, 88, 89, 98, 126, 148, 246, 258, 309, 415, 478, 491, 556, 567, 571
Ruch-Cossmann Biology Test, 436
Ruch-Popenoe General Science Test, 29-30, 424
Rugg, H. O., 187, 571
Russell, Charles, 89

S

Sampling, effect on reliability, 64-65; extent of, 64; illustration of, 63
Sangren, P. V., 39, 99, 126, 246, 258, 368, 571
Scales, quality and product, 26; and tests, 15
Scales, test-rating, 117-122; Cole-von Borgersrode, 117-120; Otis, 120-122
Scaling test items, 186
School marks, 566-571
Schorling-Sanford Plane Geometry Test, 410
Science, diagnosis and remedy in, 440-441; general outcomes, 415-420; measurable qualities, 421; tests, 29-30, 423-426, 427-432, 436-440
Scoring tests, 101, 124
Sealy, G. A., 88
Seashore, C. E., 305, 309, 475, 476, 491
Seashore Music Talent Tests, 306, 476
Sectioning of pupils, 560-561
Selected References, 8, 39, 57, 88, 98, 126, 148, 208, 245, 257, 280, 308, 347, 367, 387, 415, 442, 469, 490, 519, 540, 555, 571
Shaw, Lena, 342

Shotwell, Anna M., 39, 99, 126, 556
Sigma, *See Standard Deviation*
Silent Reading, skills, 353-355; measurement of, 356-360; remedial drills, 360-361
Simmons, E. P., 18
Simmons-Bixler Spelling Scale, 16, 18-19, 331
Sims, V. M., 53
Skinner, C. E., 285
Smith, David E., 390, 416
Smith, Dora V., 312
Smith, Homer J., 541
Smith, H. L., 39, 89, 99, 126, 246, 309, 348, 368, 387, 416, 444, 520, 541
Snedden, D., 541
Social Studies, diagnosis and remedy, 385; outcomes, 369-371; possible measurement in, 371-376; types of tests in, 376-385
Solid Geometry tests, 412-414
Sones-Harry High School Achievement Test, 551-552
Spanish, achievement tests, 462-464
Spaulding, E. R., 325
Spearman, C., 281
Spearman-Brown formula, 143
Speer, Robert K., 491
Spelling, diagnosis of disability, 335; importance of, 328; Simmons-Bixler Scale, 16, 18-19, 331, 332 ; tests, 329-335
Standard Deviation, as measure of variability, 176-182; computation of, 178-182; definition of, 176; uses of, 182
Standardization, 127-129; meaning of, 212-213
Standardized and informal tests, 20, 574-575
Standards, vs. norms, uses of, 214-215
Stanford Revision of Binet-Simon, 264
Stanton, H., 492
Statistical Problems, 159, 160, 166, 167, 170, 171, 176, 187, 200, 201, 203, 207, 244, 245
Stein, M. L., 310
Stenquist, John L., 262, 310
Stenquist Mechanical Aptitude Tests, 262
Stickney, G. E., 325
Stoddard, G. D., 39, 89, 99, 126, 208, 246, 309, 443, 541, 571
Stone, C. R., 348, 387
Strength Tests, 501-502
Strong, E. K., 300, 310
Strong Vocational Interest Blank, 300-302
Studebaker, J. W., 126, 246, 258, 415
Supervisory problems, 104
Surveys, 4
Survey tests, 26-32, 394
Symonds, P. M., 89, 310, 416, 444, 520, 556

T

Tabulation methods, 151-160
Taylor, J. S., 348
Teachers' marks, 77-78; functions of, 566-571; limitations of, 564; objectifying, 566-571
Teachers, and tests, 100-101
Teaching method, evaluation of, 562
Terman, L. M., 246, 264, 275, 281, 556; Group Test of Mental Ability, 266
Testing and the teacher, 90, 572-583
Testing programs, basis for, 103-105; steps in, 103
Test-rating scales, 117-122
Test scores, meaning of, 209-212
Tests, characteristics of, 10; equivalence of forms, 140-142; evaluation of, 105-122; kinds of, 13; meaning of, 11-12, 212, 573; related to instruction, 90; reliability of, 108-110, 142-147; responsibility for giving, 100; scales, 15; uses of, 91-97; validation of, 105-108, 129-134
Thomas, Calvin, 470
Thorndike, E. L., 38, 57, 148, 281, 319, 342, 348
Thorndike Handwriting Scale, 342
Thordike-McCall Reading Tests, 231, 233
Thurstone, L. L., 310
Tiegs, E. W., 556
Timing of tests, 123
Toops, H. A., 310
Tormey, T. J., 372, 388
Trabue, M. R., 246
Trade tests, 537-540
Travis, L. E., 316, 348
Trigonometry tests, 414
True-False tests, 83
T-scores, 185-186
Tyler, Ralph W., 57

U

Uhl, W. L., 416, 492
Ullman, B. L., 452
Ullman-Kirby Latin Comprehension Test, 452
Unit-Achievement Tests, in Algebra, 400-402; in Plane Geometry, 407-408
Unit Scales of Attainment, 237, 240, 548-550 ; samples from, 17
Upton, C. B., 416
Uses, of achievement tests, 91-96; of mental tests, 96-97

V

Validation of tests, 106-108, 130-134
Validity, defined, 105; general vs. specific, 106; in standardized tests, 130-134
Van Duzee, Roy R., 541
Van Wagenen, M. J., 16, 556
Variability, measures of, 171; Q as measure of, 173-176; range as measure of, 173; sigma as measure of, 176-182
Vocabulary tests, 362
Vocational Interest, 300-302
Vocational tests, 537-540

W

Wade, F. C., 442
Ward, C. H., 348
Warner, W. E., 541
Watkins, R. K., 443
Watts, F., 310
Weaver, C. G., 310, 542
Webb, L. W., 39, 99, 126, 556
Weidemann, C. C., 89
Weight-Height-Age Tables, 504-505
Wells, G. K., 542
West, P. V., 285, 348
Wheat, H. G., 348
When to give tests, 102-103
Whipple, G. M., 361
Whitford, W. G., 482, 492
Wiedefeld-Walther Geography Tests, 383-384
Wilkins Prognosis Test in Modern Languages, 464
Willing Composition Scale, 320
Wilson, G. M., 39, 89, 99, 126, 246, 348, 388, 520, 556, 571
Wilson Language Error Tests, 252
Winslow, L. L., 481, 490
Wise, Marjorie, 348
Witty, P. A., 337
Wood, Ben D., 89, 466, 470
Wood, Thomas D., 494, 519, 521
Wood, E. P., 85
Woody, Clifford, 39, 99, 126, 246, 571
Work-Book in Educational Measurements, 7, 106, 111, 113, 115, 124, 159, 168, 176, 200, 220, 230, 232, 239, 243, 246
Wright, J. C., 542
Wright, W. W., 89, 99, 126, 246, 309, 348, 368, 387, 416, 444, 520, 541
Writing scales, 342-343

Y

Yerkes, R. N., 310
Yes-No items, 84
Yoakam, G. A., 310, 352, 368, 388
Young, J. W. A., 416